Chapter 4

Schema-Based Segregation and Integration

In chapters 2 and 3, we have studied the two types of primitive grouping, sequential and spectral, that help to solve the problem of auditory scene analysis. We have treated them as if they were fully explainable as automatic innate processes that act without conscious control. However, this cannot be the whole story about the organization of auditory signals. In many cases, hearing a signal in a mixture of sounds depends on conscious effort and prior learning.

In the present chapter, I would like to consider the contribution made by attention and knowledge in the perceptual analysis of signals. I will not be able to provide much of an account of the internal machinery of knowledge-based analysis. Little relevant research has been done. My main emphasis will be on trying to strip away the contribution of knowledge and attention in auditory scene analysis so that the workings of the primitive process can be seen more clearly.

The nineteenth-century physicist Hermann Helmholtz distinguished between analytic and synthetic listening to complex tones. Using analytic listening, he could hear out a partial of the tone. With a more natural, synthetic attitude he would hear the holistic properties of the tone. He argued that when listening for the partials of a complex tone, "the attention of the observer has generally to be drawn to the phenomenon he has to observe. . . until he knows precisely what to look for."[509] I do not agree that segregation is the product of attention while integration is the result of inattention. However, it does seem true that attention and learning can play a role when we extract some of the components of mixtures for the purposes of pattern analysis. The goal of this chapter will be to determine whether there is any basis for distinguishing between two classes of processes in auditory scene analysis—one automatic and unlearned and the other involving learning and attention.

I have argued earlier that recognizing the timbre of a signal when it is mixed with other ones depends on scene analysis. The auditory

system must group all the components of that signal into the same perceptual stream. We saw that this was made possible by acoustic factors that bound the group of components together and at the same time distinguished them from the other ones that were present. However, J. O. Nordmark has described a number of situations in which the timbres of signals embedded in mixtures can be recognized despite the fact that there is no simple acoustic basis for segregating the partials of that signal from the mixture.[510] He has mentioned, as examples, mixtures of amplitude-modulated tones, square waves and filtered pulse trains, in which the pitch and intensity are held constant and there are no spectral or common fate cues to distinguish their partials from one another. The timbre of individual signals in the mixture is discernible. While Nordmark did not supply enough detail in his report to allow us to determine whether there was any acoustic basis for hearing out the component timbres, his examples are not the only ones that lead us to the suspicion that factors other than the ones surveyed so far are at play.

Other examples occur with speech sounds. Mixtures of two synthetic vowels can be created in which there are no spectral features that we know of that can group the harmonics or the formants into those that define the individual vowels. I have created the pair "ee" and "ah" in the laboratory. Both had the same onset and offset times, the same pitch at each moment of time, and the same loudness contour. Yet I was able to clearly hear the two individual vowels. More formal experiments using this sort of stimulus have been done by Michaël Scheffers, with similar results.[511] Scheffers fed either glottal pulses or white noise into a series of filters to create spoken or whispered vowels, respectively. In both cases, his listeners performed remarkably well in recognizing them. In the recognition test, since they had to choose the two vowels that were present in the mixture from a set of eight possibilities, the likelihood of guessing by chance was only 3.6 percent. Yet both vowels were correctly identified 45 percent of the time when the two were spoken and 26 percent of the time when they were both whispered.

There have been other cases where the auditory system has succeeded in putting spectral regions together for purposes of speech recognition despite the fact that there were acoustic cues telling it to segregate them. For example, James Cutting has synthesized syllables such as "da" by means of two formants and sent the different formants to different ears of the listener.[512] When the two formants had different fundamental frequencies, the listener heard two sounds. This occurred, presumably, because both the different spatial loca-

tions and the different fundamental frequencies indicated the existence of two separate sounds. Just the same, they heard the correct syllable (from among "ba", "da", and "ga"), an accomplishment that required them to integrate information from their two ears. Their recognition succeeded despite, rather than with the aid of, acoustic cues for grouping. A host of other researchers have found essentially the same effect—phonetic integration in the face of acoustic cues that favor the segregation of acoustic components.[513]

There is another case of auditory organization for which primitive segregation is not a sufficient explanation. The perceptual restoration of missing material in the continuity illusion does not lend itself to being explained by a scene analysis based on acoustic cues. For example, in phonemic restorations the restored sounds tend to be words that fit meaningfully into the sentence. Clearly the selection of the appropriate components from the noise burst must be based on something other than the immediately present sound because the restored material is not identical to what came either before or after the noise. It must depend not just on the sounds that are present, but the listeners' knowledge of their language.

Nature of Primitive and Schema-Based Organization

These examples all point to the fact that scene analysis can use a more sophisticated knowledge of the signal than what I have described earlier. I am therefore going to propose that there are two different processes in the construction of auditory representations, one that I will call primitive scene analysis and the other schema-driven construction of descriptions. The use of the word primitive is meant to suggest that the process is simpler, probably innate, and driven by the incoming acoustic data. The schema-driven (hypothesis-driven) process is presumed to involve the activation of stored knowledge of familiar patterns or schemas in the acoustic environment and of a search for confirming stimulation in the auditory input. This distinction is very much like the common distinction in information-processing theory between bottom-up and top-down processing.

Both processes are concerned with the decomposition of mixtures of information so that the right combination of information can enter into the description of an environmental sound. However, the primitive mechanism does this without reference to the recognition of specific familiar sounds, whereas the sophisticated ones select the right components as a part of the process of matching stored schemas for familiar environmental sounds to the incoming data.

Properties That May Distinguish the Two Systems

The approach that I plan to take is to assume that the primitive segregation process employs neither voluntary attention nor past learning. Therefore I will call the process schema-driven if it uses either of these capacities. In my definition of the schema-driven process, I may be accused of conflating two different processes. It is possible that the use of voluntary attention and the use of prior learning are not the same thing at all. However, I have two reasons for grouping them. First is, I am trying to strip away everything that is not part of the primitive process, to see whether some "pure" properties of the primitive process can be isolated. I feel that it is necessary to do so because there exist, in the research on perceptual segregation, some contradictions that can be resolved only by assuming that there is more than one process at work. Second, another way of describing voluntary attention is to call it programmed attention, or attention that is under the control of an inner process that is trying to find some particular pattern in our sensory input. Since I look at schemas as control systems that deal with patterns in the environment, it is natural to think of them as the ones that govern voluntary attention.

We know that both attention and prior learning can affect the process of extracting signals from mixtures. Let us consider these in turn and see how they function.

The concept of attention encompasses two distinct facts about the human mind. The first is that there are processes that can select part of the currently available sensory information for more detailed processing. Let me give an example: When you are asked to pay attention to sensations arising from your left foot, you will suddenly become aware of a set of experiences that were previously not part of your consciousness. This selective role of attention has been studied in audition mainly in a situation where a listener is trying to pay attention to one spoken message in the presence of a second one.[514] Many factors have been shown to affect the ability to do this. Among these are the acoustic differences between the two messages. For example, if the messages are spoken on very different pitches, as happens when one is spoken by a man and the other by a woman, it is easier for the listener to attend to one of these and not be interfered with by the other. Other differences, such as in location, can assist this segregation. The key factor that makes us think of this process as attention is that the listener is *trying* to hear one of the two messages. This notion of trying is central to the definition of attention.

Why, then, did I argue earlier that these experiments tell us about primitive perceptual grouping? Why are they not really experiments on attention? After all, the listener is usually trying to accomplish the

task of hearing a subset of sounds as a separable pattern in a mixture. This is a difficult question. When we use a person as a subject in an experiment, the successful completion of the task involves a large constellation of capacities, including the ability to hear the signal, to understand the instructions, to sit still, to attend to the signals, to make some sort of judgment, to turn the judgment into words or numbers, and so on. If this is so, how do we know which process has been affected when the experimenter has manipulated some variable in an experiment? The answer is that we do not ever really know for sure, but we can make reasonable guesses based on common sense and on our prior understanding of these processes. For example, if it is easier to distinguish two voices when they are at different pitches, it is hard to see how the difference in pitches might affect the ability of the listener to sit still in the chair; we are not familiar with any mechanism whereby this ability could be affected by such a difference. On the other hand, if we attribute the influence of acoustic differences to their effect on our ability to pay attention to messages, this makes sense to us for two reasons: First, it agrees with our own experience outside the laboratory, and second, we can make up a plausible mechanistic story about how it could occur.

However it is more difficult to know whether to attribute the result of an experiment to attentional processes or to primitive processes that act independently of attention, because we do not have a very good understanding of either class of process. Therefore we cannot decide whether the effects of different stimulus variables are more consistent with what we know about one or the other. Perhaps one clue to the nature of the effect is the effect of *trying*. If trying harder makes a task possible, we will attribute the improvement to attention because this effect is consistent with our beliefs about attention. Unfortunately, we have no such touchstone for the primitive scene-analysis process. Suppose we found that some variable affected segregation independently of how hard a listener was trying to achieve segregation; are we to attribute this effect to a property of attention or to that of a preattentive process?

Apart from the role of effort there are other signs by which we recognize the presence of attention. One is that we have a more detailed awareness of things that are the objects of attention than of things that are not. This has led to descriptions of attention as a process that employs more of the mental resources of the listener than would otherwise be allocated to analyzing a sound.

Another sign is that we have trouble paying attention to too many things at the same time. This has led to the argument that there is a limited pool of resources that attention can make use of.

A third observation is that as we become highly practiced in a task, it comes to require less attention. For example, a person who is learning to drive feels swamped by all the things that must be attended to at once. However after a number of years, the driving process can become so automatic that the driver can drive while lost in thought about other matters. This has led to the view that attention is involved in the novel coordination of a number of separate skills. Although each skill, in itself, might already be highly practiced, the coordination is not; attention is no longer required when that coordination becomes a highly practiced skill in its own right. This view of attention implies that we cannot tell whether attention is or is not involved in a task by simply knowing what the task is. We have to know what resources are being employed by the person who is doing it.

We are left with a view in which attention is seen as an effortful process that coordinates existing schemas in a new task and in which the coordinating process (presumably some newly assembled schema) can control only a limited set of mental resources. The intimate connection between attention and learned skills (schemas) is one reason that I want to bundle attention and learning in a common package and distinguish them, as a pair, from primitive segregation.

An example of the role of attention in scene analysis is the intense concentration that it sometimes takes to be a good subject in an experiment in which you have to hear out one component of a mixture in which the acoustic conditions favor fusion. You must fix the sound of what you are listening for strongly in your mind and then try to hear it in the mixture. I find that in such cases if my attention is distracted by even a slight extraneous noise, I am no longer sure whether I can hear the target sound or not. I also find that I get better at it with practice. This improvement is not the mark of an automatic innate process.

The role of learning in the perceptual organization of sounds is easily demonstrated. In one condition of an unpublished experiment that I did many years ago, the listeners had to listen for a simple tune that covered a wide range of frequencies, and ignore other notes that were present in the same frequency range as the tune.[515] Under some circumstances the notes of the target tune were made much louder than the distracting tones. In this condition the listeners could hear the tune. But when the tones of the tune were of the same loudness as the distractors, most listeners could no longer hear the tune. In the course of running this experiment, I undoubtedly heard the tune thousands of times. Eventually I found that I could hear the tune even when it was no louder than the distractors. It was not as if I could

avoid hearing the distractors as part of the stream, but I knew which tones were melody and which were distractors (since the distractors always were the same and situated in the same positions in the melody) and I could somehow mentally bracket the distractors and hear the tune. Jay Dowling also found that a familiar tune could be more easily extracted from interfering sounds than an unfamiliar one.[516] The role of learning can also be seen in the continuity illusion. For example, it is possible to obtain illusory continuity of a tune through an interrupting noise. However, this effect is stronger when the tune is more familiar.[517]

The characterization that I have given to primitive and schema-driven scene analysis has distinguished them in terms of the psychological mechanisms that are presumed to underlie them. It is also necessary to distinguish them by their effects on the process of stream formation.

First let us consider the primitive process that has been the subject of the preceding chapters. Its role is to employ heuristics that have been formed in the phylogenetic evolution of our sensory systems and to put together auditory features that have probably come from the same source. As clues to the correct grouping of features, it uses acoustic properties that tend to be valid in a wide variety of auditory environments, without regard for the specific meaning of the sounds. That is, the clues are the same whether we are listening to music, to speech, or to cars moving past us on the street. They include such things as frequency proximity, spectral similarity, correlations of changes in acoustic properties. These clues are the ones that form the basis for the so-called Gestalt principles of grouping. The Gestalt psychologists, mostly describing visual perception, argued that these principles were the expression of innate processes. They used two forms of evidence to support this argument. The first was the fact that even newborn animals showed these forms of perceptual organization. The second was the observation that, even in adult human beings, camouflage that was based on these basic principles of grouping could prevent us from recognizing even highly familiar shapes.

The schema-based process, on the other hand, makes use of knowledge about specific domains of experience. Before going into details, I want to clarify my use of the word "schema." This word is used by cognitive psychologists to refer to some control system in the human brain that is sensitive to some frequently occurring pattern, either in the environment, in ourselves, or in how the two interact. Many psychologists have speculated on the form in which the knowledge about such regularities is packaged in the brain, and the functional, and maybe even anatomical, structure that holds this knowledge in

the brain has been given different names, such as cognitive structure, scheme, schema, frame, and ideal.[518] They all describe a computing structure that controls how we deal with one particular regularity in our environment. Such structures can be either very concrete in nature, such as the coordination of perception and movement patterns that is required to tie shoelaces, or extremely abstract, such as the coordination of grammatical and lexical patterns required to form a sentence. Often a convenient way to label these schemes is by the environmental regularity that they deal with. For example, the Swiss psychologist Jean Piaget studied the schema for causality.

When we perceptually analyze our auditory input, we can make use of schemas about recurring patterns in the world of sound. These patterns vary enormously in their degree of complexity and abstraction. My knowledge about the tonal glide that forms the song of the cardinal is both simple and concrete. Knowledge about the properties of a familiar word is more abstract because the word can appear in quite different acoustic forms when it is spoken by different persons or with different intonation patterns or speeds. Knowledge about a grammatical form (such as the passive) is even more abstract, because it is defined not in direct acoustic terms, and not even in terms of patterns of specific words, but in terms of the patterns formed by classes of words (nouns, verbs, auxiliaries, and the like). Yet patterns at all these different levels of abstractness are thought to be dealt with by schemas in the human mind.

Often the sound that we are listening to contains more than one pattern. If each pattern activates a schema, there will be a combination or pattern of schemas active at that time. Sometimes the pattern of activated schemas will form a larger pattern that the perceiver has experienced in the past. In this case the pattern of activity of schemes can evoke a higher-order schema. This occurs, for example, when individual items are recognized as different types of words, such as nouns or verbs, and their arrangement is recognized as a sentence.

There is considerable argument in psychology about whether any schemas can be innate (the nativism-empiricism controversy). We know that a vast number of schemas are learned. As examples we can cite the coordinations involved in all motor skills, the particular patterns used by our own language to convey different meanings, and the social patterns viewed as polite in our culture. The fact that these are learned is not disputed because they vary across individuals and cultures. The argument in psychology is mostly concerned with the complexity of the schemas that can be innate. Not even the most extreme nativists would deny that many schemas can be learned because they are aware of examples like the ones I have given. Similarly,

no empiricist will deny that certain things are innate, such as learning ability itself or the tendency to extract or to group certain sensory inputs. The argument lies in how complex the innate schemas are in the human being. For example, are there ready-made ones that exist for the human face, for the sexual act, or for the basic structure of human language?

For the purposes of this book, it would not be necessary to decide such issues if we thought that schemas, whether learned or innate, worked in similar ways. I think that they do. One strong reason for thinking so is that they must collaborate smoothly.

Among the jobs that a schema must perform is to apply itself appropriately to a situation. This is what Piaget referred to as assimilation. Computer scientists refer to it as pattern recognition. Each schema must have its own particular methods of evaluating a sensory input to determine whether the pattern that it cares about is there. In the domain of sound, this evaluation has an inherent time dimension: The evaluation processes look for temporal patterns.

Schemas are scene-analysis processes by their very nature. Why? Because it is in the nature of a schema to analyze the sensory data for evidence for the presence of some pattern. Since it knows what the temporal patterning of the evidence should be, it can use this knowledge to extract the pattern from a mixture. If we grant that this is so, we are tempted to ask why we should have primitive scene analysis processes at all. We might argue that schema-driven scene analysis is more powerful because as we become more and more familiar with a pattern and build a more detailed schema for it, we can do better and better in extracting it from mixtures.

We can answer this challenge by remembering that schemas can only do the scene-analysis job on familiar patterns. Yet if we are to learn about patterns in the first place, so as to make them familiar by forming schemas for them, we need some primitive processes that are capable of extracting them from their acoustic contexts. Although the primitive processes probably always make a useful contribution to scene analysis, it is in the realm of unfamiliar patterns that they are absolutely essential.

Still, schema-driven analysis of acoustic input can be very powerful. One reason for this is that patterns extend over time. If four or five elements of a temporally unfolding pattern have already been detected (such as the first several notes of the national anthem), the schema enters a state in which it is primed to detect later elements. In this musical example it would find a note (and the listener would experience it) even if it were entirely obliterated by noise. We have already examined this phenomenon and others like it in our discus-

sion of the continuity illusion. Notice that I have referred to this technique as powerful and not hallucinatory. The power resides in this: Because the evidence for the national anthem is of such high acoustic quality prior to and after the obliterating noise, the weakening of the criteria so that we may accept the obliterated sound as the missing note is quite well justified. By gathering evidence over an extended period of time, we are protected from damage to the quality of evidence at particular moments.

Some very simple examples of schema-based segregation can be found in the research of Charles S. Watson and his colleagues at the Central Institute for the Deaf in St. Louis. Suppose we play a sequence of several tones to a listener twice, but change the frequency of one of the tones (call it the target) on the second presentation, and test for the listener's accuracy in detecting the change. It is usually found that the accuracy can be up to 10 times worse than it would have been if the target and its repetition had been the only tones presented.[519] Apparently the tone becomes embedded in the sequence, and its own particular properties do not draw the attention of the listener. However, the accuracy goes up dramatically if the same sequence is used over and over again and the same position in the pattern is designated as the target on every trial, instead of using a new sequence and target position on each trial.[520] It also becomes easier when the target is marked out by increasing its intensity by 15 dB relative to the rest of the tones,[521] or by making its duration longer than that of its neighbors.[522] In all these cases, the listener is able to come to treat the relevant tone as different from the rest of the sequence. Yet the segregation is difficult, not usable by every listener, and benefits from training. I would argue that this segregation is based on a different mechanism than the one that would segregate it if it were in a wholly different frequency range from its neighbors. The latter mechanism produces a large and automatic perceptual isolation.[523] I believe that it is schema-governed attention that allows us to focus on the slightly longer or louder tone. A schema that describes the sequence can be learned and then used to select the target. This, I believe, is the mechanism for the analytic listening that Helmholtz described in 1859.

Even when schema-based recognition processes become strong, as when an adult listens for a familiar pattern, the primitive processes still play a role. If a schema-based integration could become entirely independent of primitive scene analysis, then camouflage would be impossible. Yet it happens all the time. We can make it hard for a listener to hear a familiar sound (such as a tune) by embedding it in a pattern of tones (even softer ones) in the same frequency region.

If the activation of a schema were designed to be independent of primitive scene analysis, it would have to occur even when the supporting evidence was not packaged in a single object. Schemas would therefore be very susceptible to errors resulting from inappropriate grouping. Parts of two people's voices could form a word that neither had spoken.

Our best guess at the present time is that the human brain has some method of combining the benefits provided by the two systems. We do not know whether we give more weight to the grouping decisions provided by one system or the other. Probably this varies with circumstances. It may be that the more practiced a schema-governed process is, the more weight is given to it in relation to the primitive scene-analysis process. This may be why speech perception can do two things that seem to violate the assumption that primitive scene analysis is dominant. One is that it can hear speech sounds even when it has to combine information across primitive groupings of sense data to do so. This achievement has been referred to as duplex perception and it will be the topic of discussion in chapter 7. Another thing that it can do is to hear two vowels out of a mixture even when there is no primitive acoustic basis for segregating the spectrum.

How Do We Know They Should Be Distinguished?

We have given some reasons for wanting to distinguish between a primitive and a schema-driven process, but that is not enough. We have to see whether there is empirical evidence for this distinction.

The first thing to mention is that there is confirmation that the primitive processes are unlearned. In chapter 1 we discussed the research of Laurent Demany who showed that infants aged $1\frac{1}{2}$ to $3\frac{1}{2}$ months of age showed evidence of auditory stream segregation.[524] Unfortunately, this is the only study that I know of on this important topic, probably because it is very difficult to do research with young infants. We can only hope that more studies will be forthcoming.

We are justified in distinguishing the two forms of scene analysis if we can distinguish two different patterns of causality in the evidence. One technique for doing this is to hypothesize that there are two different causal nexes and see whether this assumption allows observations to fall into simpler patterns.

In looking for a primitive and a schema-governed mechanism, we must remember that when we set a human being a task, the response that we observe is the result of the activity of the whole complex human being. When we do experiments, we have to hope that the variance attributable to the factor that we manipulate is due to its

effects on only one mechanism, in this case, the primitive or the schema-based system for auditory scene analysis.

This approach is made difficult by the fact that the same sorts of acoustic properties that serve to segregate sounds through Gestalt-like processes of organization can also be the bases for recognizing these sounds. Therefore they appear in schemas. We can have a stored representation of a sequence of sounds that encodes the temporal pattern of its elements and their relative pitches; otherwise we would not be able to remember music. Our memory complex for changes in timbre over time is what enables us to distinguish one instrument from another. Yet we know that these same factors, pitch, timing, and timbre form the basis for primitive grouping as well. This makes it very hard to distinguish primitive from schema-based segregation. We can, nevertheless, make the attempt.

As an example let us look at the perceptual segregation of rapidly alternating high and low tones. Two factors that affect the degree of segregation are the frequency separation between the high and low tones and the rate of the sequence. However, Van Noorden has showed that the rates and frequency separations at which one gets segregation depend on the intention of the listener.[525] His findings were shown in figure 2.2 of chapter 2. The upper curve, the "temporal coherence boundary," showed the boundary between integration and segregation when the listener was trying to achieve integration. The lower curve, the "fission boundary" showed what happened when the listener was trying to segregate the streams. We saw that if the intention of the listener was to hear two distinct streams the frequency separation that was required was only about three or four semitones. Furthermore, this requirement was almost independent of the rate of the tones. This contrasted strongly with the case in which the listener was trying to hold the sequence together. Here not only did the separation have to be greater, but it depended strongly on the rate of the tones.

The dependence of the boundary on the frequency separation and rate of the tones in one case and not in the other is a clue that two different separation processes were at work. Why should frequency separation not affect the fission boundary? Remember that in this task the listener was trying to hear out one of the streams from the mixture and that furthermore the stream was very simple, the periodic repetition of a single tone. This meant that the listener could store a mental description of the sound and its periodicity and try to match that stored description to the sequence. This was obviously a case of schema-based attention, and it was very successful in analyzing the sequence. Since it did not depend on the relation between the target

tones and the other ones, it was not sensitive to the frequency/time spacing between these two subsets. The lower limit of three or four semitones for the fission boundary, beyond which the listener could no longer pick up the tones of one stream without the intrusion of the others, may represent some physiological limitation on the sharpness with which our auditory attention can be tuned. It may not be a coincidence that the value of three or four semitones is about the same as the width of the critical band, the region within which a tone played simultaneously with another one will mask it.[526]

On the other hand, the temporal coherence boundary, where the listeners are trying to hold the sequence together, shows a strong effect of the proximity of the high and low tones to one another. A trade-off of frequency and temporal separations is evident; the faster the sequence, the less the listeners can tolerate a frequency separation between the tones. The organizational process opposes the conscious intentions of the listeners to deploy their attention in a certain way. Surely the up-and-down oscillation pattern is not an unfamiliar one to the listeners. Therefore a schema exists for that pattern. But the primitive organization into two streams opposes its application to the sense data.

A property that might be useful in distinguishing the two types of segregation is symmetry. When we cause a sequence of sounds to be segregated by increasing the frequency separation of subsets of tones, this improves the listener's ability to listen to either the higher or the lower sounds. The effect is symmetrical for high and low tones. This may be the mark of primitive segregation. When the segregation is by timbre, the same symmetry is found. If the difference in the brightness of the timbre of two interleaved sequences of tones is increased, both sequences become easier to isolate. If their spatial separation is increased, again both become easier to hear as separate sequences.

However, I do not think that the same thing occurs when the segregation is based on a schema for a familiar sound. As we increase the difference in familiarity between two sound patterns that are mixed together by making ourselves more and more familiar with just one of them, although we may become increasingly more skillful at pulling the familiar one out of the mixture, this does not, in itself make it easier to hear the unfamiliar one as a coherent sequence. Dowling did experiments in which the notes of two melodies were interleaved.[527] Some of them looked at the effects of pre-familiarizing his listeners with an arbitrary sequence of sounds that subsequently was mixed with another sequence. The listeners were able to detect a familiar sequence more easily than an unfamiliar one. However, in one ex-

periment he prefamiliarized the listeners not with the target sequence itself but with another one that was to serve as the interfering background. After familiarizing his listeners with this background sequence, he interleaved another one with it and asked his listeners which of two possible sequences the new one had been. He found that familiarity with the background did not assist his subjects in isolating the target melody.

Here is another informal observation. In the course of running an experiment, I was obliged to listen to a melody that was interleaved with distractors in the same frequency range. Naive subjects could not hear the melody, but after hundreds of exposures I reached a point where I could. The distractors in this sequence never varied, so I became very familiar with them, too. The familiarity with the overall sequence of tones, melody tones interleaved with distractors, allowed me to hear out a part of it as the melody. The remaining tones were heard as adornments to the melody.

My ability to use my familiarity with the background contrasts with Dowling's results. In Dowling's case the listeners were prefamiliarized with the background taken alone, whereas I became familiar with target tones and background at once. I do not think I was segregating them in any primitive perceptual way, but was coming to deal with a primitively coherent sequence by using a schema to divide it into two parts. Dowling's listeners could not do this because they had a prior schema for only the background tones that they had been practiced on, not for the two parts together. We can conclude that having a schema for a background pattern does not help us to integrate a target pattern. A schema helps only if it contains either the target pattern itself, or both the target and background sounds.

It can be argued that the role of the primitive segregation processes is to *partition* the input, while the job of the schema-governed process is to *select* an array of data that meets certain criteria.

The existing experimental results on stream segregation by loudness differences may not be due to primitive segregation. They may be the result of a deliberate attempt on the part of subjects to focus their attention on the less intense or more intense set of sounds. I would remind the reader of my earlier discussion of research on loudness-based streaming in chapter 2. The conclusion was that the results of the research did not show the kind of symmetry of segregation that I am using as the hallmark of primitive segregation. This does not mean, of course, that primitive streaming by loudness does not exist, only that it has not been demonstrated yet.

The factor of speed may also help to distinguish the two kinds of constraints on stream segregation. In many, if not all, cases of primi-

tive grouping, speed serves to increase the segregation based on acoustic factors such as frequency separation. On the other hand, when the segregation is based on the recognition of a familiar subsequence, the segregation may worsen with speed. This may happen because speeding up a familiar pattern distorts it or because the process of directing one's attention towards the sequence of components in turn may be less effective at higher rates.

The effects of speed on schema-based integration of sequences might be found if we were to do an experiment that used a task that has been employed by Diana Deutsch to study the recognition of melodies. Deutsch did some experiments in which simple familiar folk tunes were distorted by randomly deciding for each tone, independently, whether to move it up exactly one octave, move it down an octave, or leave it where it was.[528] Moving a tone by an octave keeps the chroma (note name) the same. That is, a C-sharp is still a C-sharp. However, scattering the tones into three octaves destroys the size of the changes in pitch height (roughly speaking, the change in frequency of the fundamentals of successive tones). It appears that pitch height relations are more fundamental than chroma relations in creating the perceptual organization that allows us to recognize a tune. Therefore the listeners could not recognize the distorted tunes. However, if they were told in advance what the tune was, they could hear it. Under these conditions it appeared to Deutsch that the listeners were mentally generating the tone sequence, abstracting the chroma of each note in turn, and verifying it against the corresponding tone in the auditory input.

My expectations about this task would be as follows: Recognition could be accomplished only at fairly slow rates of presentation, and it would not be achievable at six tones per second (whereas the original tune would undoubtedly be recognizable at this rate). At higher rates, recognition would be impossible for two reasons. The first is that primitive grouping by pitch height would increase at higher speeds, creating compelling patterns that were unfamiliar. Second, the complex control of attention and of testing hypotheses against the input would be too slow to keep up. The task that we have just discussed is a rather complex one, undoubtedly more intricate than the task of recognizing a familiar pattern of sounds in a mixture. Just the same, it would indicate that complex processes of directing ones attention to the individual elements of a sequence may be rate limited.

Another possible experiment would use a sequence in which a familiar tune was interleaved with distractor tones in the same frequency region. Assuming that at a low rate the listener could detect the familiar sequence in the mixture, we would expect that speeding

up the sequence would cause primitive grouping by frequency to destroy the tune as a perceptual unit.

Because studies comparing the effects of speed on primitive and knowledge-based partitioning of signals have not yet been done, our expectations can only be conjectural.

I think that there is another difference between primitive and schema-governed segregation. Their temporal scope is different. The acoustic relations to which the schema-based processes are sensitive can span longer time intervals than those that the primitive processes look at. For example, in the phenomenon of phonemic restoration, when a sound in the middle of a word is obliterated by a noise, the word that the listener hears is one that fits appropriately into the framework of discussion and the current topic of discussion. This could depend on words that had arrived much earlier in the discussion. We already know that the subject of the immediate sentence can affect the restoration of words. Take, for example, the sentence "The *eel was on the ———", where the asterisk represents a noise burst and the blank represents the word that is in the last position. If the final word is "orange" the listeners tend to hear "*eel" as "peel", if the final word is "axle" they are more disposed to hear "wheel", and if the final word is "table" they favor "meal".[529]

Here the critical last word was very close in time to the target word (perhaps a second later). However, it is possible that it might be effective at a much greater distance if it preceded the critical word. Imagine a series of sentence such as this: "I have had a lot of trouble with the right rear wheel of my car. It seems to be very loose. It worried me a lot, so since I had a free day last Sunday I decided to do something about it. I looked at the *eel very carefully." Here, the topic of discussion is set up two sentences before the one with the missing phoneme, perhaps 5 or 6 seconds earlier. Yet I would expect to find a preference for restoring "wheel".

The effects of most acoustic variables on primitive scene analysis are much more local. If we alternate a sequence of high and low tones more slowly than two tones per second, no compelling segregation by frequency will occur. A 500-msec separation between tones of the same class seems to place them beyond the range of primitive interaction. There seems to be one exception to this time limit for primitive grouping. If you alternate high and low tone repeatedly at perhaps 10 tones per second, the strength of segregation builds up over a period of at least 4 seconds and perhaps longer. If the rate of alternation is slower, the preference for segregation can take as much as a minute to build up.[530] However, in this case, the hundredth high tone is not grouping with the first one but with the ninety-ninth and the

hundred-and-first. The long exposure biases this grouping, but the grouping is still between nearby sounds.

In summary, we hope to be able to distinguish between primitive and schema-based scene analysis by the role of effort, the need for learning, the symmetry of the segregation, the interaction with speed, the time span over which they group material, and, in general, the different causal pattern observed in different tasks. With luck, the distinction between the two types of mechanisms will help us to understand some apparently discrepant results in the research literature.

Does Learning Affect Streaming?

Schemas can be acquired through learning. Therefore schema-driven scene analysis should be affected by learning. What evidence is there in the research literature about the effects of practice on our ability to integrate a stream of sounds and segregate it from a mixture?

Let us return to the research of Dowling on pairs of interleaved melodies.[531] We have seen that the subjects could not make use of their knowledge of the background melody to segregate the foreground melody if the two were in the same frequency range. But they could benefit from even verbal naming of the target melody itself. In one condition, he gave them the name of the melody and then played the mixture several times to them. Recognition was successful after an average of three to four presentations. Their ability to use the melody's name shows that familiarity can produce segregation in the absence of primitive segregation, but the fact that they required a number of exposures to the mixture even after they were told the name and that they detected the tune only with considerable effort suggests that the segregation process was not automatic, but that a schema-driven attentional process was involved.

Do Regular Patterns Form More Coherent Streams?

Cognitive psychologists think of schemas as descriptions of regular properties of the environment. It is natural, therefore, to ask whether regular auditory patterns form more coherent streams than irregular ones do. If the process that is responsible for the formation of streams is one that makes predictions about the properties of the next sound, regular patterns should afford better prediction, and therefore better integration.

Jones' Rhythmic Theory of Attention The role of certain kinds of schemas on the organization of sounds has been described by

Jones.[532] I have already briefly described the theory in chapter 2. Her research has concerned itself with the recognition and memorization of short melodic and rhythmic patterns. It has given evidence for the fact that human attention can be deployed in a rhythmic manner. According to Jones' "rhythmic theory" of attention, as we listen to a pattern of sounds, the process of attention is capable of anticipating the position of the next sound on the time dimension, the pitch dimension, and others as well. It therefore can prepare itself to pick up the next sound and connect it with the preceding parts of the sequence. A predictable sequence allows its components to be caught in the net of attention, while unpredictable elements may be lost.

Rhythmic attention, as described by Jones, is restricted to the use of certain kinds of rules; they describe the incoming sequence as a hierarchy of embedded units. A hierarchy, in the sense meant here, is formed when units serve as the parts of larger units and the latter, in turn, serve as the components of yet larger units. In her theory, any two sequential units that are grouped together to form higher-order units must have a particular type of relationship to one another. The later unit must be formed from the earlier one by repeating it after imposing a relatively simple transformation on it. This type of relationship is often found in music where repetition with variation is a common architectural principle. The theory describes rules for the encoding of pitch patterns and rhythmic patterns by rules of this type.

It is assumed that the attentional system tries to apply the rules on the fly to an incoming signal. If the rules do not apply, either because the sequence is irregular or because suitable rules have not yet been formed, some of the sounds may be lost from attention and not incorporated in the ongoing stream.

Jones offers an explanation of why a sequence of tones alternating between two frequency regions will tend to form separate streams if the alternation is rapid but not if it is slower. Her explanation, like that of van Noorden's,[533] depends on the assumption that attention cannot shift too rapidly along a sensory dimension. I have discussed the limitations of this explanation in chapter 2.

This theory sees auditory streaming as the result of the operation of a temporally extended process of attention. My own preference is to distinguish the schema-governed attentional process from primitive processes of grouping. In this dichotomy I see Jones' theory as weighted heavily toward the schema-driven process. It does not distinguish primitive from schema-driven bases for organization.

The theory has led to a number of studies, the results of which seem to suggest that a rule-based account of this type (which does not distinguish between primitive Gestalt-like grouping and grouping

based on learned patterns) can be used to explain auditory organization. There are certain considerations that argue that the apparent success in explaining sequential integration by means of hierarchical pattern schemas cannot be taken at face value.

One is that the results of the experiments are often explained equally well by a primitive grouping mechanism as by a schema-governed attentional one. The reason for this is that the aspect of Jones' theory that predicts stream segregation, deriving from the idea that attention cannot shift too rapidly from one value of an auditory dimension to another, generally makes the same predictions about segregation that a Gestalt-like grouping account, based on proximities in frequency and time, would do.

In some experiments done by Jones and her collaborators, there are observed groupings of sounds that cannot be explained by her theory but are explainable by primitive organizational processes. For example, in one study by Jones, Maser, and Kidd, a subset of four middle-range pitches was found to form their own group despite the fact that the formal sequence-describing rules did not describe any particularly strong relation between them.[534]

There is another reason for not accepting the results at face value. While the theory proposes that it is perceptual integration that is assisted by the schemas, it has actually been tested most often with memory tasks, which are very sensitive to the activity of schemas. Therefore the results can be interpreted as effects on memory, not on perception. The listeners have been asked either to recall what they have heard (using some sort of written notation) or to compare two sequences that have been presented one after the other.[535] Cognitive psychologists have frequently described memorizing as the use of preexisting schemas to incorporate the study material or as forming new schemas that are appropriate to the material. Schemas, then, are thought to be intimately involved in the memorizing process. Therefore when a theory proposes that it is explaining a perceptual phenomenon, such as auditory streaming, by means of a process that involves the coding of the stimulus through pattern rules, it has to be careful not to use a task that is heavily dependent on memory or on conscious strategies for finding patterns by which the material can be more easily remembered. It is possible to find ways of minimizing the contribution of memory, for example by allowing the subjects to make their decisions about the stimulus while it is still present, by making the decision task as simple and overlearned as possible, or by using some measure other than description of the pattern (for example, asking how many streams there are or what the rhythm is).

Although the theory describes the use that attention makes of rhythmic schemas of the repetition-with-variation type, it does not describe temporal schemas of any other type. Yet many of the auditory patterns that we must deal with in daily life are not repetitious (or rhythmic) in this way. The theory seems to have been created in order to deal with the type of patterns found in music in which there are hierarchies formed out of discrete tones. I do not see how it could deal with the irregular time patterns that are found in many natural situations, as, for example, in the unique sound of an automobile braking to a stop or the modulating sound of a train whistle, or the long drawn out intonation pattern of the word "sure" that communicates irony. Yet we track such sounds over time and form schemas that describe their temporal properties.

A final point is that the theory does not address spectral organizational processes at all. Even if the theory is intended to account only for sequential integration, we know from the evidence that I presented earlier that there is a strong competition between spectral grouping and sequential grouping and therefore the accounts cannot be divorced.

In assessing the role of schemas in auditory scene analysis, however, it would not be wise to focus on the more specific statements of a particular theory. Instead it would be better to address some general assertions that any theory of Jones' type is likely to produce. First let me say what I think this type is. The class of theory that is involved is one that assumes that the organization of an auditory sequence into streams is not a consequence of the primitive process that I have described in earlier chapters, but is one that employs sequence-prediction rules. Certain statements follow from this assumption:

1. Learning will affect stream segregation.
2. Regular patterns will be found to form more coherent streams.
3. Streams are created by attention and by the search for regular patterns in the stimulus sequence.

The following sections will attempt to assess the truth of these assertions in the light of the available experimental evidence.

Any theory that bases the perceptual integration of a series of tones on the capacity of an attentional process to anticipate the position of the next tone in frequency and in time would seem to predict that a random sequence of tones (unpredictable by definition) can never be coherent since no rule can predict the properties of the next tone. However, I have listened to sequences of tones that occurred at a

regular temporal spacing but at random frequencies. As these sequences are speeded up, groups of tones that happen to be near one another in frequency seem to pop out of the sequence and form a brief stream of their own. Since the sequence is random, these accidentally occurring groupings come and go in different frequency regions.

Van Noorden did a related experiment that bears on this issue.[536] He wanted to see whether the average frequency difference between successive tones would still predict the perceptual coherence of a sequence of tones when the observer did not know what the next tone would be. He constructed sequences in the following manner. Two decisions were made before the sequence began, a frequency for the initial tone (I) and an average frequency difference (D) between successive tones for that particular sequence. The first tone was at frequency I. Every subsequent tone was either higher or lower than its antecedent by either the frequency separation D itself, or by D plus or minus one semitone. The choice between higher or lower and between the three possible frequency separations was made at random. This kind of sequence can be called a "random walk" and if completely random can walk right outside the frequency range that the experimenter is interested in. So an upper and lower frequency boundary was imposed and when the randomly chosen frequency increment for the next tone would have moved it outside one of these boundaries, another one of the possible frequency changes was randomly chosen instead. On each run of the experiment, the average absolute deviation in semitones from one tone to the next was set at a particular value. The listeners adjusted the speed of the sequence to the fastest at which a single coherent stream could still be heard with no tones detaching themselves into a separate stream.

Van Noorden reported that at a very fast speed (20 tones per second) and at a small average frequency separation (one semitone), the sequence sounded like a ripply continuous tone because under these circumstances the separation is below the fission boundary. If the same sequence was played more slowly (about 80 msec per tone) the separate tones were heard but the sequence was still continuous in quality. At the other extreme, if the average frequency separation was 25 semitones, and the speed was high, no temporal coherence could be heard except for a few tones here and there that were not temporally adjacent but happened to be close in frequency to one another. Over a set of sequences with different frequency separations, as the average separation became larger, the speed of the sequence had to be slowed down more to maintain the experience of a single coherent sequence. The results closely resembled those obtained in experiments where high and low tones alternated in a regular manner.

When these results were plotted on top of results from the regular sequences, treating the average frequency separation in this experiment as analogous to the fixed frequency separation of the other experiments, the curves were quite similar. Van Noorden concluded that "previous knowledge of what tones are coming does not influence the temporal coherence boundary, and that listening for temporal coherence is probably not a question of moving the attentional 'filter' to and fro actively but of following the tone sequence passively." It is this passive process that I have called primitive scene analysis, and contrasted with schema-directed scene analysis.

Van Noorden's experiment does not prove that a prediction-forming process is *never* involved in auditory stream segregation. It merely shows that frequency separation is sufficient in itself to promote segregation without the participation of a rule-extrapolating process. Van Noorden's failure to find a difference between experiments that did and did not permit prediction of the next tone may derive from the particular nature of the tone sequences.

Does the Auditory System Track and Project Trajectories?

Let us now go on to ask whether there is any evidence that regular patterns are easier to integrate than irregular patterns and, if they are, what this implies about the importance of prediction in perceptual integration and segregation.

The first kind of regular pattern that we should look at is the regular trajectory. This is a smoothly ascending or descending change in frequency (or pitch) with time. In my discussion I will equate frequency with pitch, because in the experiments that have been done with trajectories, the two were never varied independently. In some cases, as well as being smooth, the change has been quantitatively regular: Each step in pitch was the same size on some plausible scale, such as the raw frequency scale, or the diatonic or the chromatic musical scale. If the trajectory were drawn on a graph whose dimensions were time and the chosen scale, it would always look like a straight line. Sometimes the trajectory was formed from a sequence of steady tones and sometimes it was a continuous glide.

According to any sequence-prediction theory of perceptual integration, an acoustic sequence of this type should be easily heard as an integrated whole. This follows because the rule that generates them is exceedingly simple. However, the coherence of simple trajectories might be expected on other grounds as well. The Gestalt approach to perceptual organization would explain the coherence of a tonal trajectory through the concept of good continuation. In visual displays, straight lines and smooth curves have this property.

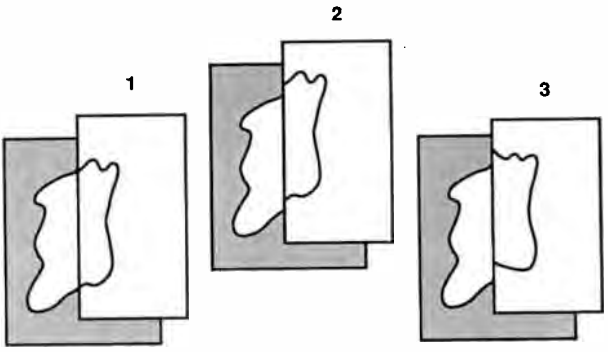


Figure 4.1
Left: an irregular figure that continues smoothly through a vertical line and is perceived as a unitary form. Middle: a discontinuity is created by displacing the right side vertically at the line, but the slopes of the lines on the two sides of the division are equal. Right: the slopes are not the same on the two sides of the line.

A visual example of the effects of smoothness and discontinuity on perceptual integration is depicted in figure 4.1. Panel 1 shows an irregular figure that passes smoothly through a vertical line that separates two surfaces; the continuity causes the figure to be seen as unitary, overcoming the fact that the two halves are drawn on backgrounds that make them appear to be on separate planes. In panel 2, the right side is displaced upward and the figure no longer looks unitary; its halves seem to lie on the separate planes. In panel 3, the contours on the two sides of the vertical line are made not to match in slope as they do in panel 2. This further dissociates the two halves. Despite the fact that there is no simple rule that can describe the shape of the curved figure, the continuity still has its effects.

In contrast with sequence-predicting theories, the Gestalt approach would not see this preference for smooth change as necessarily due to a process that tried to generate a rule to describe the whole shape. Instead, the preference would be based on the smoothness of transition of each part of the form to the next.

We can see that the integration of smoothly changing auditory figures, such as a smoothly ascending sequence, might be expected both from a sequence-prediction mechanism or from a primitive integration mechanism that preferred sequences that had low levels of discontinuity.

Evidence That the System Does Not Project Trajectories I will start reviewing the research on the subject of trajectories by describing some

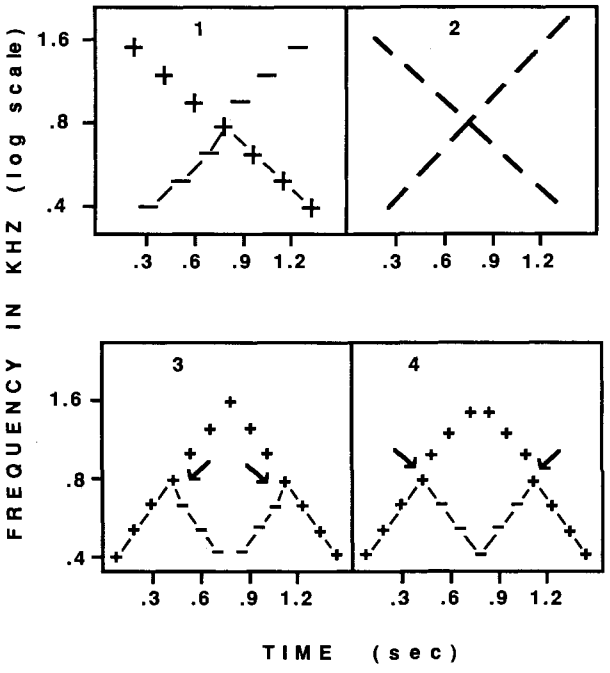


Figure 4.2
Competition between segregation by frequency region (the broken connecting lines) and by the following of a trajectory (pluses). 1: the crossing of an ascending and a descending pattern of tones. 2: crossing trajectories in which the individual tones are short glides and overlap in time. 3: the rhythm favors the trajectory. 4: the rhythm favors segregation by frequency range. (Adapted from Tougas and Bregman 1985a.)

experiments that failed to find evidence that the auditory system extrapolates trajectories.

A very common observation is that when an ascending trajectory of tones intersects a descending trajectory, as in part 1 of figure 4.2, it is very difficult to hear either the ascending or the descending trajectory crossing the other one. If you focus on the descending one, you tend to follow it down to the intersection, and then your attention is forced to shift to the second half of the ascending glide so that what you hear is the V-pattern that is formed by the higher tones. Alternatively you may hear the two lower halves as a pattern that ascends to the point of intersection and then descends again. The auditory system prefers to group sequences that are in the same frequency region than to follow a trajectory. The effect is compelling and is almost impossible to overcome by a voluntary attempt to follow one of the

trajectories. A large number of experiments have studied how we perceive such crossing patterns and they have all come up with the same result.[537] The effect has been exploited in Diana Deutsch's scale illusion in which the tendency to group tones that stay in the same region can produce a failure to group all those tones that fall on a common trajectory or that come to the same ear.[538]

An experiment by Yves Tougas and myself tried in a number of ways to get the listeners to hear the trajectories as crossing.[539] One way was by the use of rhythm. Parts 3 and 4 of figure 4.2 show a simplified pattern: The pluses represent a rising and falling trajectory and the minuses show other tones that could group by frequency with the first part of the trajectory. When they did, the listeners heard two streams, one consisting of the tones joined by the short lines and the other consisting of the remaining pluses. These streams united tones that were in nonoverlapping frequency regions. The difference between parts 3 and 4 is that in the former, if the listeners group the tones that form the long trajectory (the pluses), they are selecting tones that are equally spaced in time and have a regular rhythm. However, when they group tones that are in the same frequency region (those joined by the lines), they are selecting tones that have an uneven rhythm. The arrows point to intertone intervals that are half the duration of the others. It might be imagined that the factor of rhythmic regularity would favor the integration of the trajectory, but it did not. In stimuli resembling those of part 4, the time intervals were rearranged so that homogeneity of rhythm favored grouping by frequency region, yet this made no difference. The stimuli of parts 3 and 4 were heard in the same way. The overwhelming organization was by frequency proximity with either rhythm.

Tougas and I tried to improve the likelihood of the formation of the trajectory-based organization by means of two other factors, shown in part 2 of the figure. We converted the tones into short glides that were aligned on the trajectory. We also introduced a temporal overlap of the tones of the ascending and descending glides so that if the auditory system were to group tones that were in the same frequency region, it would have to either backtrack in time, skip an interval of time, or break the unity of a tone. Neither of these manipulations of the pattern succeeded in causing the trajectory to be favored. There was only one manipulation that caused the grouping to be based on the trajectory. We took a pattern similar to the one in part 1 of the figure and enriched the timbre of the ascending pattern of tones. This succeeded in causing the listeners to hear two trajectories, one rising and the other falling. However, it is misleading to interpret these results as revealing a trajectory-based organization. Since we

know from other research that tones that are different in timbre may form separate streams, there is no need in this experiment to attribute any influence at all to the trajectory.[540]

Another failure to find a trajectory effect in auditory grouping was obtained in an experiment done by Howard Steiger and myself.[541] This experiment was reported earlier in this volume and its stimuli illustrated in figure 2.17 of chapter 2. It employed a repeating cycle in which a pure-tone glide (A), acting as a captor, alternated with a complex glide BC containing two pure-tone glide components, B and C. It was possible for the captor to capture one of these components (B) out of the complex glide into a sequential stream consisting of two successive glides, A and B. This tended to happen when A and B had similar slopes and center frequencies. In one condition, A and B were aligned on a common trajectory. According to any theory that says that sequential grouping was the result of a prediction process, this arrangement should increase the likelihood that A would capture B, since the position of the beginning of C is predictable from the slope of A. However, no such increase was found.

This result is incompatible with what Valter Ciocca and I found when we studied the illusory continuity of glides behind a loud noise burst.[542] As reported in a later section of this chapter, the Ciocca-Bregman experiment found that the perceptual restoration of the missing portion of the glide was better when the glide segments that entered and exited from the noise were aligned on a common trajectory. However, there were many differences between the two experiments and it is hard to say which of them was critical. The tasks were different, capturing a component from a complex glide versus restoring a glide in noise. The durations of the glides were different, those in the perceptual restoration experiment being much longer (500 msec versus 130 msec). It is tempting to think that the fact that the experiment with the longer glides showed the trajectory effect means that it takes some time for the auditory system to measure the slope of a glide. However, we should recall that in the capturing experiment, despite the shortness of its glides, there was evidence that the auditory system could measure slopes. The capturing of B by A was best when they both had the same slope. Therefore we know that the slopes were registered. However it did not help to have the glides aligned on a common trajectory.

Another difference between the experiments was that the time interval between the end of the first glide and the onset of the second was different, only 20 msec in the capturing experiment and from 150 to 200 msec in the perceptual restoration experiment. Yet, contrary to what might be expected under the assumption that it is harder to

extrapolate across a longer interval, it was the experiment with the longer interval that showed the trajectory effect. Maybe the longer intervals, together with the longer glides, permitted the listeners to use a schema-governed attentional process to group the sounds.

A final difference is that the interval was silent in one case and filled by noise in the other. Perhaps noise and silence give different instructions to the auditory system. Perhaps a silence tells it that the sound is now ended and that it can now terminate the following of the glide's slope, whereas a loud noise onset does not cause it to terminate the following of the slope in this way. With so many differences, we cannot really say why the results were different, but their difference casts doubt on the unqualified assertion that a primitive grouping process tracks trajectories.

Another failure to find a trajectory-extrapolating effect occurred in an experiment by Gary Dannenbring at McGill.[543] His stimuli were illustrated in figure 1.15 of chapter 1. A connected sequence of alternately rising and falling glides (linear on a frequency-by-time scale) was presented. In one condition, the peaks (where the ascending glide turned into a descending glide) were replaced by bursts of white noise. The listeners perceptually restored this turnaround point and experienced the glides as alternately rising and falling. To measure how the restored glide was experienced, they were asked to match the highest pitch that they heard in a second glide pattern (in which there was no interruption by noise) to the highest pitch that they heard behind the noise. These matches showed that the highest pitch that the listeners actually heard was not the one that the deleted peak would have had but was approximately at the highest pitch that the glide had reached before the noise (actually it was even a bit lower). Therefore the listeners did not extrapolate the pitch change into the noise, but simply interpolated between the pitches on the two sides of the noise.

Here is an informal observation of another failure of capturing by trajectory. The reader may recall the experiment that Alexander Rudnicky and I reported on the capturing of tones by a stream of tones of the same frequency.[544] The stimuli are illustrated in figure 1.7 of chapter 1. The listener's task was to recognize the order of the two highest tones (A and B). This was made hard by the presence of flanking tones (labeled F). However, the flanking tones could be captured by a series of captor tones (C) that was near in frequency to them. This made the order of tones A and B easier to recognize. When Rudnicky and I saw these results we wondered whether we could "tilt" the pattern and still get the same effect. We created the pattern that is illustrated in figure 4.3, hoping that the ascending

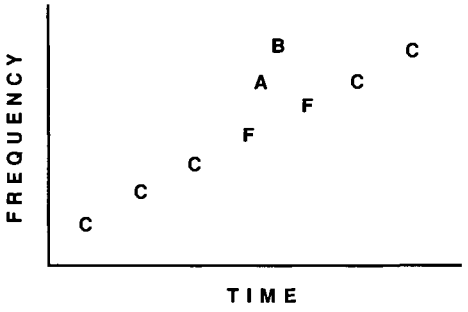


Figure 4.3
A modification of the stimulus pattern of Bregman and Rudnicky (1975) in which the capturing tones form an ascending sequence.

trajectory formed by the C tones would again capture the F tones so that they no longer interfered with the independent perception of A and B. However, we could find no evidence that this occurred and consequently abandoned the experiment.

Yet another failure to find a trajectory effect was found in an experiment that tried to capture a harmonic out of a speech sound. Researchers in Christopher Darwin's laboratory had done a number of experiments that showed that it is possible to capture a spectral component out of one of the formants of a short vowel and alter the perceived quality of the vowel. They had been able to do this by introducing, just before the vowel, a sequence of tones that were of the same frequency as the target component; however, they found that when they introduced, before and after the vowel, a sequence of tones that was aligned with the target partial to form a rising or falling sequence, the capturing showed no effect of the trajectory itself, but was entirely explainable in terms of the proximity of the frequencies of the nearer tones to that of the target.[545]

If a regular trajectory is a basis for the primitive grouping of sounds at all, it appears to be a rather weak one, easily overpowered by other factors. In chapter 2 we saw how the onset portion of a continuous long glide could be captured out of the total glide by being preceded by a short glide starting at about the same frequency. When this happened, the short glides and the initial portions of the long ones formed a stream of their own leaving the remainder of the long glides behind as a residual stream. The pattern was illustrated in figure 2.19 of chapter 2. Surely nothing can form a more regular trajectory than a glide that can be drawn as a straight line on log-frequency-by-time

axes. Yet this did not prevent its being broken up by a frequency-based grouping.

The results of these experiments have led us to look for another reason for the findings of Dannenbring and myself, reported in chapter 2, and illustrated in figure 2.24.[546] In one condition of this experiment, the segregation of alternating high and low tones in a rapid sequence was reduced when the end of each tone consisted of a brief frequency glide pointing toward the frequency of the next tone. When we published this result we attributed it to a principle of grouping that makes use of the predictability of the position of the next sound (in a frequency-by-time space) from the final transition of the previous one. However, the results may not have been due to this sort of process at all. The presence of frequency transitions might simply have reduced the frequency separation between the end of one tone and the beginning of the next and the observed effect might have been entirely attributable to the principle of grouping by frequency proximity. If this were true, and if the time separation between the tones were to be reduced by speeding up the presentation rate of the sequence, splitting would be expected to eventually occur since increasing the presentation rate increases the segregation of sequences of tones that differ in frequency. We did observe this in informal observations.

We did some informal experiments in which we moved the tones closer together in frequency but pointed the transitions away from the adjacent tone. If the adjacent tone was lower in frequency, the transition went upward and vice versa. This arrangement did not reduce the integration of the sequence in comparison with the case in which transitions pointed to the next tone as long as the end point in the frequency transition was at the same frequency separation from the onset of the next tone in both cases. In fact, there was even a hint that the pointing-away transitions improved the integration. However, since this result did not accord with what we believed at the time to be a genuine trajectory pointing effect, we did not pursue this research to the point of publication.

Evidence That the System Projects Trajectories The preceding discussion has described a number of studies that failed to find that the auditory system extrapolated trajectories. However, this is not the end of the story. I am now obliged to review a number of research findings that seem to support the existence of an extrapolation mechanism. I will argue that there is an alternative explanation in every case.

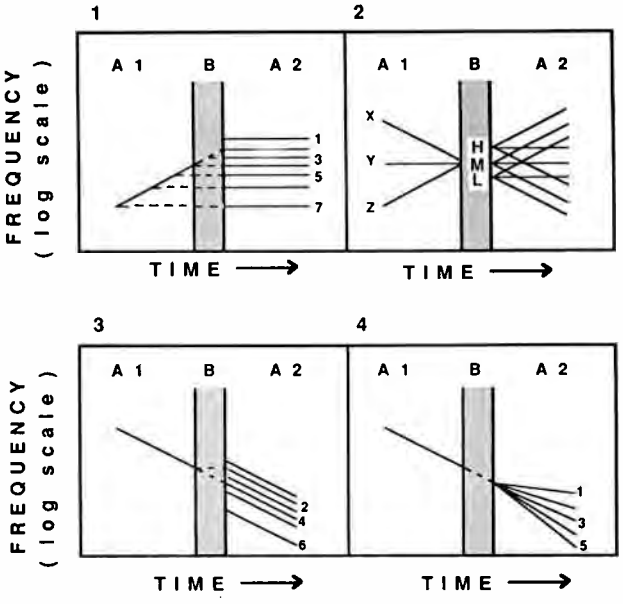


Figure 4.4
The alignment of glides and their apparent continuity. (From Ciocca and Bregman 1987.)

Trajectory-Based Integration of Streams Among the studies that seem to support a trajectory mechanism are a group reported recently by Valter Ciocca and myself on the continuity illusion.[547] Earlier Gary Dannenbring had shown that listeners could perceptually restore the continuity of a repeating pattern of ascending and descending pure-tone glides when a portion of the pattern was replaced by a noise burst.[548] However, Dannenbring never varied the alignment of the audible glide portions along a common trajectory, so there was no way to tell whether this alignment played a role in the continuity illusion. Ciocca and I used the type of stimulus illustrated in figure 4.4. It had three parts: a glide (A1) that preceded the noise, the noise itself (B), and a glide (A2) that followed the noise. A number of experiments were carried out in which the slopes of A1 and A2 were varied as well as the point in frequency at which they abutted the noise.

The listeners were pretrained on stimuli very like the ones in the main experiment except that in half of them, the glide actually continued underneath the noise at an audible level. They were trained to judge how sure they were that the glide continued through the noise.

In the actual experiment the listeners were led to believe that the glide continued through the noise but it never did. They were nonetheless asked to use a nine-point scale to rate how clearly they could hear the glide under the noise.

The figure illustrates the stimuli in a number of the experiments. Panel 1 shows an ascending glide (A1) followed by a noise burst, which in turn is followed by a set of possible steady-state tones (A2). One of these A2 sounds, the second from the top, began at just that point in frequency that the A1 glide would have reached by the end of the noise if it had been continuing underneath the noise. This predicted point of the exit of the trajectory from the noise is given the name "trajectory point" or T. The results showed that the A2 that gave the strongest impression of continuity was not the one at the T point, but the one that was at the highest frequency that A1 had reached before the noise. The more the other A2 glides departed from this best position, the less strongly they favored the perception of continuity.

If this experiment were taken alone, we would have to rule out the idea that the auditory system measures and extrapolates trajectories. However, others that we did seem to offer a different picture. The stimuli for the second experiment are shown in panel 2 of the figure. There were three versions of A1, ascending, descending, and steady, and there were nine versions of A2. Each A1 was combined with every A2 to create 27 different stimuli. One factor that distinguished the different A2 glides was whether or not they started at the trajectory point; the other was whether they were parallel to the A1 glide on a log-frequency scale. For each A1 there was one A2 that was exactly aligned with it on a log-frequency-by-time trajectory.

The results showed that for the falling A1 glide, the exactly aligned A2 glide was the best facilitator of continuity. The second best was the A2 glide that was the exact reversal of A1. This latter glide began at exactly the frequency that A1 had left off and rose to the frequency at which A1 had started. We can call it the "mirror-image glide". The other A2 glides did not promote perceptual restoration as well. In the conditions where A1 was rising, the rank of these two A2 glides was reversed. The mirror-image A2 glide was best and the trajectory-aligned A2 glide was second best. Finally, when A1 was a steady-state tone, the trajectory-aligned and the mirror-image A2 glides degenerate, by definition, into a common form, namely the A2 tone that is at the same frequency as A1. This A2 promoted continuity the best when A1 was steady-state.

Generally speaking, when A1 and A2 had the same slope, a trajectory effect was observed, the best continuity being obtained when

they were aligned on a common trajectory. The continuity decreased as we moved the starting point of A2 away from the trajectory point. However, when the slopes of A1 and A2 did not match, the best continuity was obtained when the final frequency of A1 and the initial frequency of A2 were the same.

One way to explain the success of the mirror-image glide and the role of the similarity of the final part of A1 and the initial part of A2 in promoting continuity is by assuming that there is a preference for having the A2 glide as close as possible in frequency to the A1 glide. This is an example of the principle of frequency proximity. Presumably the nearest parts of A1 and A2 would have the greatest weight in this computation of proximity, but the superiority of the mirror-image glide to steady-state glides supports the idea that the more distant parts of the glides also count in the proximity calculation, the mirror-image A2 glide being best because it occupies exactly the same frequency region as does the A1 glide.

These results can be taken as further examples of the pervasive influence of frequency proximity in the sequential integration of streams. What is new is the apparent trajectory affect that seemed to be present whenever A1 and A2 had the same slope.

We went on to study this trajectory effect using stimuli that are illustrated in panel 3 of figure 4.4. In this experiment A2 always had the same slope as A1, but it could start at the trajectory point or above or below it. The A2 glides that started at or near the trajectory point (glides 3, 4, and 5) produced good continuity. For those that were nearer in frequency to A1 (glides 1 and 2) continuity was much worse.

Panel 4 shows the stimuli for a further experiment in which all the A2 glides began at the trajectory point but had different slopes. Of these, those that had the same slope or a more extreme one (glides 3, 4, and 5) gave good continuity while those with shallower slopes (1 and 2) did not.

The requirement of having A1 and A2 lie on a common trajectory was not precise. The exactly aligned glides were always good, but not always best. However this outcome is a reasonable confirmation of the occurrence of sequential grouping based on the alignment of trajectories. I can think of a possible explanation for the absence of more precise effects. We have no reason to believe that the log-frequency-by-time coordinates on which the trajectories can be described as straight lines are the ones that the auditory system uses to match up trajectories. The fact that we obtained any trajectory results at all suggest that this coordinate system is not too far off, but we do not know how far that may be.

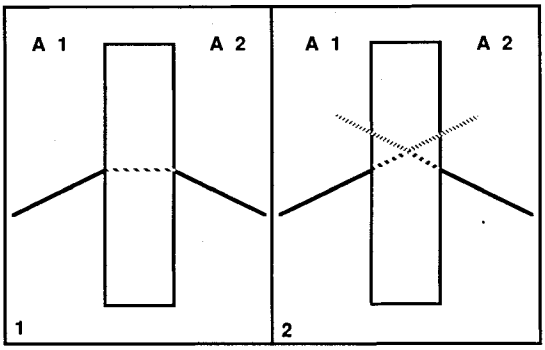


Figure 4.5
Illustration of two possible mechanisms for the restoration of glides through noise: interpolation (1) and extrapolation (2).

Clearly, this experiment supports the existence of a trajectory-extrapolating mechanism in perceptual restoration. However, we have to be very careful in interpreting it. There are two issues in the restoration of glides: One is the nature of the cues that are used to decide whether A1 and A2 are parts of the same glide. The present experiment supports the idea that A1 and A2 tend to be grouped whenever A2 has the same slope as A1 and when it exits from the noise at the trajectory point.

The second issue is the mechanism that is used to calculate what trajectory should be *experienced* during the noise. We could imagine the two mechanisms illustrated in figure 4.5. The first one, shown in panel 1 of the figure, works by interpolating a glide between the last frequency of A1 and the first frequency of A2. This mechanism treats restoration as analogous to stretching a rubber band between the last frequency estimate before the noise burst and the first one that follows it. The alternative mechanism for restoration, shown in panel 2, treats it as the extrapolation of a glide along a trajectory that is partially missing. The mechanism might be viewed as doing something analogous to placing a straight ruler along A1 and drawing a straight line forward past its end, then placing the ruler along A2 and drawing a line backward past its beginning until it reaches the line drawn from A1, then restoring, as the missing trajectory, the part of each line that does not extend past the other.

The experiments done by Ciocca and Bregman do not distinguish between these two alternatives. Although they show that alignment on a common trajectory may be the cue that promotes the grouping of A1 and A2, we could imagine either mechanism for restoration

going along with this cue. It seems, no doubt, to go most naturally with a restoration mechanism that involves extrapolation of trajectories (the ruler method), since both the use of the cue for grouping and the method of restoration require the auditory system to extrapolate trajectories. The interpolation (rubber band) method of restoration seems, on the other hand, to go together most naturally with the frequency-proximity cue for integration, since neither requires the system to be able to extrapolate along trajectories. Even so, it is possible that the decision as to whether A1 and A2 are parts of the same sound is based on how well their trajectories lined up, but that the rubber band method is used to generate the experienced pitch trajectory.

The experiments by Ciocca and Bregman could not distinguish between different possible restoration mechanisms because they never measured *what* the listeners heard, but only *how clearly* they heard it. The only attempt to measure it was in an experiment by Dannenbring that was described earlier in this chapter. When the listeners restored the peaks in a rising and falling glide sequence they did not extrapolate the pitch to the missing peak but acted as though the auditory restoration process was stretching a rubber band between the nearest audible portions of the glides.

We must remember that the Ciocca-Bregman research showed that if glides are not aligned on a trajectory, the best matching is between mirror-image pairs. This mirror-image relation was, of course, the one that caused the grouping of the glides in the Dannenbring experiment. Perhaps it could be argued that if the glides are grouped by the frequency-proximity rule, the restoration will be by the rubber-band mechanism, and it is only when glides are grouped by their alignment on a trajectory that the ruler (extrapolation) mechanism will be used. However, the realization soon sinks in that in the case of aligned glides, the two proposed mechanisms will always give identical results. Therefore we have no need to ever postulate the existence of an extrapolation mechanism in the restoration of glides.

The only reason for believing in it is that the more complex forms of restoration seem to involve some sort of extrapolation. For example, we can restore the part of a word that has been deleted by noise, and this is obviously not achievable by an interpolation between the edges of the remaining portions of the word. However, this is not an acoustical extrapolation. Stored schemas for words are involved. The audible portions of the word are brought into contact with these schemas and one of these word schemas, selected by its acoustic match to the evidence and by its appropriateness to the semantic context, is perceptually restored. This is far from a simple process of

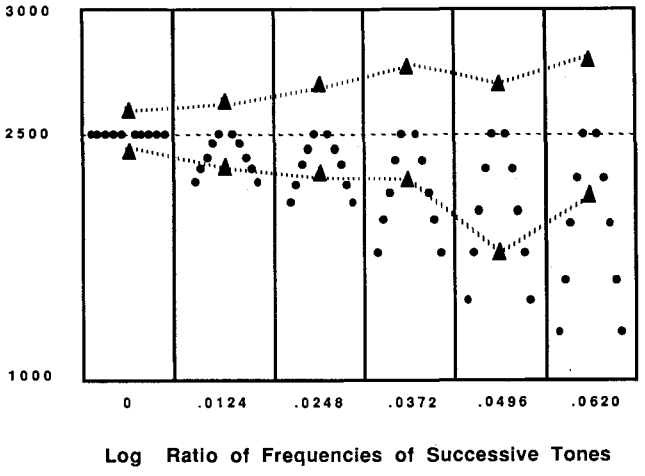


Figure 4.6
Six of the tone patterns used by Heise and Miller (1951). Circles: tones in the pattern. Upper triangles and lower triangles: mean values for the upper and lower limits for the adjustable tone.

extrapolation of an ongoing pattern and it clearly involves the use of stored schemas.

My best guess, then, is that glide trajectories are restored by interpolation unless a higher-order schema is responsible for the restoration. In the latter case, extrapolation can occur. Having brought schemas into the picture, however, it is tempting to ask whether even the trajectory-alignment criterion for whether to treat A1 and A2 as parts of the same glide derives from the learning of schemas. In other words, did the subjects in Ciocca and Bregman's experiments group the frequency aligned glides better because they had learned schemas concerning continuous forms of change? The data from this experiment cannot answer this question. However, we should remember that the stimulus was slow, and the listeners were making an effort to hear the sound. These factors would favor the use of schemas. It is even possible that the preference for the mirror-image continuation was based on a schema and was not just the result of frequency proximity.

An early experiment done by George Heise and George Miller has generally been interpreted as showing that trajectories can be extrapolated by the mechanism that creates auditory streams.[549] They presented listeners with a sequence of 11 pure tones, played at the rate of eight tones per second. Six of their sequences are shown by the

circles in figure 4.6. These patterns form straight lines or inverted V's on log-frequency-by-time coordinates. The listeners could control the frequency of the sixth tone in the sequence (the middle tone). In each trial, a particular sequence was played over and over without pause while the listener adjusted the tone. If its frequency was set either too high or too low, it seemed to pop out of the sequence. This meant that it had been perceptually segregated from the sequence. The triangles superimposed on each pattern in the figure are the highest and lowest frequencies to which the listeners had to adjust the tone to get it to segregate. When the patterns were straight lines, as in panel 1, we cannot tell whether the "best position" for the tone is based on interpolation or extrapolation, since both mechanisms would give the same result, so let us examine the inverted V patterns. If we focus on the highest point of adjustment (the upper triangles) it appears that they fall on the point in frequency at which the trajectory would have reached a peak. Furthermore as the trajectories get steeper, the highest point gets higher as if trying to fall at the predicted highest point. This is the evidence for extrapolation.

However, you should not look only at the highest-limit adjustments but at the lowest-limit ones as well (lower triangles). The algebraic mean of the upward and downward adjustments made by the subject can be taken as an indication of the neutral point, the place where the tone fits in best. This value was not at the projected vertex. Instead it was at a point that was quite close in frequency to the tones that came before and after it, and in no case was this neutral point extrapolated outside the range bracketed by those tones. This suggests that the auditory system did not project the trajectory to the apex of the pattern but rather preferred a tone that lay within the frequency range of the pattern. But what of the tendency for the highest-limit setting to be higher when the trajectories were steeper? An examination of the figure shows that the lowest-limit results show a comparable effect. They get lower when the glides get steeper. This cannot be explained by the assumption that the best location for the adjustable tone is moving to a higher frequency. Instead, what appears to be happening is that the predicted point for the middle tone is getting less definite as the trajectories get steeper. That is, as they get steeper, they accept a wider range of tones as fitting into the pattern. A similar effect was found in the Ciocca-Bregman experiment that was discussed earlier.

An important fact in the results of Heise and Miller was that the tones following the adjustable one, and not just those preceding it, affected segregation. The best position for the adjustable tone in the ascending straight line patterns is higher than for the inverted V pat-

tern despite the fact that the tones preceding the adjustable tone are identical in the two cases. This is because the later tones are higher in the straight line patterns.

To summarize, the results of Heise and Miller provide no evidence for a frequency-extrapolating mechanism that "looks for" a tone at the next predicted frequency. All of their results can be explained by the tendency to prefer a tone that lies as close as possible in frequency to the nearest neighbors on both sides of it, with perhaps a weak effect of its proximity to next-to-nearest neighbors.

Their results are different from those by van Noorden that we will describe later and will illustrate in figure 4.7. Perhaps the difference lies in the experience of the listeners. While those of Heise and Miller were graduate students in psychology, van Noorden served as his own subject at a time when he had been listening extensively to such patterns. It is possible that van Noorden formed schemas for the patterns and these were the source of the trajectory effect.

Order of Unidirectional Sequences Is Easier to Report The experiments that we have surveyed so far in looking for a trajectory effect have used a number of methods that were presumed to be sensitive to the streaming of tones. The ones I plan to look at now use a different method, the report of the order of tones in a sequence. They all show that the order is easier to report in smoothly ascending or descending sequences (glissandi) than in irregular ones. However, I plan to argue that these effects are not due to a trajectory-based principle of primitive grouping but to different factors.

Results Explainable by Other Factors The first factor that may be responsible for a number of these effects is memory. A sequence that continuously ascends or descends in pitch is probably easy to remember because of the simplicity of the rule by which it can be remembered.

An experiment by Nickerson and Freeman in 1974 found that the order of a sequence of sounds was easier to report when it followed a unidirectional trajectory.[550] However, as I have shown elsewhere in this volume, the detailed structure of their results shows that they were probably not due to stream segregation at all.

Warren and Byrnes also showed that it was much easier to report the order of sounds in repeating glissandi (of four 200-msec tones) than in irregularly ordered tone cycles.[551] However, the task that showed this effect most strongly was one where the listeners had to report their order verbally. This task is heavily dependent on memory. However, a second task was also used in which the listeners took

four cards, each representing one of the sounds, and arranged them on the table to report the order of the sounds. This task reduces the memory load because the listeners can break up the task of reporting the entire sequence into one of relating each sound to the one that precedes it and inserting its card in the proper place. They need not remember the entire sequence at once. When the task was changed from verbal reporting to this card-ordering method, the superiority of glissandi over irregular patterns dropped from 110 percent to 21 percent. This suggests a strong contribution from short-term memory to the observed effects.

Another piece of evidence that the results were not due to primitive grouping was that the effect of frequency separation in this experiment was in a direction opposite to its known effects on primitive grouping: a greater frequency separation made order identification easier. If the difficulty of the irregular patterns came from their breaking into substreams, the task should have become harder when an increase in the frequency separation created more segregated streams. Another hint that higher-level cognitive processes rather than primitive grouping were responsible was that the effects occurred at a rate of presentation of five tones per second, a slow rate that, according to van Noorden, requires at least a 15-semitone separation between adjacent tones before compulsory segregation occurs.[552] None of the conditions of Warren and Byrnes employed such large frequency separations.

These observations suggest that the effect could be related to the problem of building a mental description of the event sequence, rather than simply grouping the events at a preattentive level. Therefore this experimental demonstration of the ease of reporting glissandi cannot be taken as evidence that there is a primitive process that tracks a trajectory and prefers to integrate its members into a single stream.

Stephen Handel and his associates at the University of Tennessee have found effects of trajectories in experiments on auditory patterns conceptualized in the framework of Gestalt psychology. McNally and Handel experimented with the effects of different patterns on streaming, using repeating cycles of four sounds, presented at the rate of five sounds per second.[553] Their listeners reported the order of the sounds by using a card-ordering task. One type of pattern was made by arranging the order of four tones separated by 100 Hz (750, 850, 950, and 1,050 Hz). Regular trajectories (four in a row ascending or else descending) were easier for the subjects than were irregular orders of the four tones. Although this result seems to show that the auditory system is extrapolating trajectories following a regular

trajectory, there is an alternative explanation. The speeds and frequency separations that gave the reported effect with McNally and Handel's four-tone pattern suggest that stream segregation was not a compelling percept at all. The greatest separation (300 Hz) in this sequence was about six semitones. As I mentioned in relation to the Warren-Byrnes experiment, van Noorden's studies show that at five tones per second you would have to have about a 15-semitone separation to obtain compelling segregation. The separation between the highest and lowest tones used in McNally and Handel's experiment was closer to the separation at which van Noorden obtained compulsory integration, and the separations between the nearer pairs of tones were even closer to this value. So although the ability of the subjects to report the order of the tones may have been affected by an irregular order, the irregularity probably had its effects through another mechanism. I have suggested that this mechanism is memory.

In a different experiment, Handel, Weaver, and Lawson used the same rate of five tones per second but employed greater frequency separations.[554] In one of their sequences, four tones were spaced by octaves (750, 1,500, 3,000, and 6,000 Hz). On each listening trial, a particular arrangement of these four tones was continuously repeated for a minute while the subjects made their judgments. Two kinds of judgments were required. They indicated how the tones seemed to group perceptually by drawing lines to connect the tones in a picture that was similar to a musical score. On other trials, they were not told the order of the tones but had to judge it by arranging four cards on a table. In sequences in which the four tones were equally spaced in time, those orders that formed an ascending (1234) or a descending (4321) glissando were perceived as more coherent than other orders. That is, they were more often judged to be a single sequence and their order was more often judged correctly. But there is some discrepant data. Some of the irregular orders were also judged as forming a single stream on a substantial proportion of trials, up to 47 percent of the time in some conditions. This probably happened because of the slowness of the sequences. Again, because of the fact that the conditions were not those that led to a compelling stream segregation, we cannot rule out the participation of a more complex cognitive process, such as memory, in producing these results.

Although the slow rate in these studies makes their interpretation questionable, Pierre Divenyi and Ira Hirsh have observed superior performance on glissandi with rapid sequences of tones (1 to 30 msec per tone) where stream segregation would have been expected to occur.[555] The order of rapid unidirectional sequences of three tones has also proved easier to report than other orders of these tones.[556]

However, even if some or all of the superiority of glissandi in these studies really did result from a preattentive grouping process, the effect can be explained without resorting to an extrapolation of trajectories. The frequency-proximity principle alone would be sufficient to explain their results. I will offer the argument that in a glissando, some of the competition of groupings that are found in up-and-down sequences will not be present. This lack of competition may be responsible for the coherence of glissandi.

We should remember that frequency proximities in sequences of tones always compete with one another.[557] The grouping of temporally nonadjacent tones, as in the streaming phenomenon, occurs because these tones are closer to each other in frequency than adjacent tones are, and, in addition, are not very far apart in time. In a repeating glissando, however, there is never any tone that is closer to a given tone (A) in frequency than its sequential neighbors are, except for the next occurrence of tone A itself on the following cycle. Since a glissando is likely to include at least three or four tones, the successive occurrences of tone A itself are likely to be far enough apart in time to prevent them from forming their own stream. Hence A is most likely to group with its nearest temporal neighbors. This argument deduces the coherence of glissandi as a special case of the principle of frequency proximity under the assumption that proximities compete with one another.

There is, however, an observation by Leo van Noorden that cannot be explained by the competition of proximities. One type of stimulus that he made was an eight-tone sequence, presented at the rate of 10 tones per second and covering a range of one octave. The listener heard this sequence cycled repeatedly.[558] Several different patterns were used. Four of them are shown in figure 4.7. The difference between the two top patterns is that although the first high tone is equally far in frequency from its nearest temporal neighbors in both patterns, in the second pattern, the transitions before and after it seem to point to it. It is on a trajectory (roughly speaking) that is formed by the first three tones. When listening to these two patterns van Noorden found that the first high tone remained part of the sequence in the second stimulus, where the pointing was present, though not in the first. (The second high tone was perceptually segregated from the sequence in both cases.) The two patterns in the bottom panels of the figure were heard as integrated sequences. He explained this by saying that reversals in the direction of pitch change are infrequent in this sort of pattern. This explanation amounts to saying that the auditory system prefers to hear a pattern that continues a frequency change. It is as if, in a sequence of tones ABC, the frequency relation

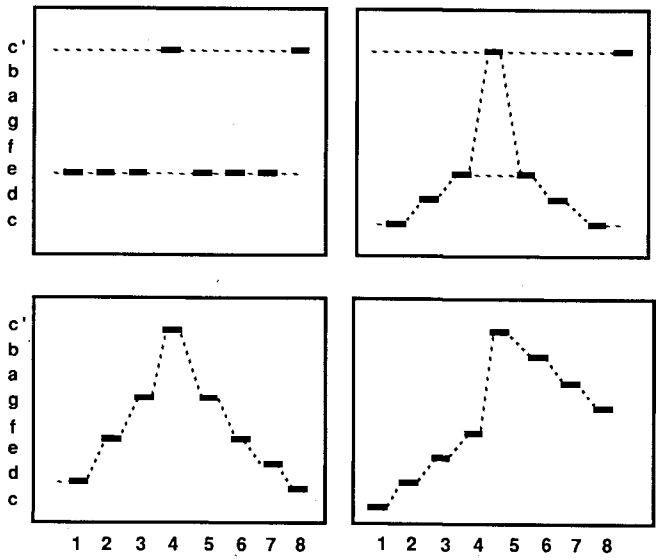


Figure 4.7
Four patterns used by van Noorden (1975) to study the effects of continuity of frequency change on grouping.

between A and B becomes part of the description of the pair and C is more readily integrated into the sequence if the relation BC is like the relation AB. If the relation between AB and BC is different, a drawing of this sequence would show a "corner" at B. Corners may tend to disrupt the integration of sequences. This, of course, is quite different from saying that the auditory system is making up a predictive schema for the sequence or is tracking a trajectory. Instead, it puts the source of the effect at a more primitive level of the auditory system, a process that forms a discontinuity when it detects something like a corner. This is reminiscent of the Gestalt principle of continuity that I described earlier and illustrated in figure 4.1. Guilford and his coworkers in the 1930s found that if listeners had to listen to two sequences of tones and judge whether the second was the same as the first, if the change was made by altering a tone at the apex of a V so as to decrease the sharpness of the V, it was usually not noticed, whereas if it increased the sharpness it was noticed more frequently.[559] This observation also suggests the existence of a system that detects corners and makes them more salient.

Whatever the process is that groups tones that fall on a common trajectory, it has different effects than the one that groups those that are near to one another in frequency. Let me illustrate this with a

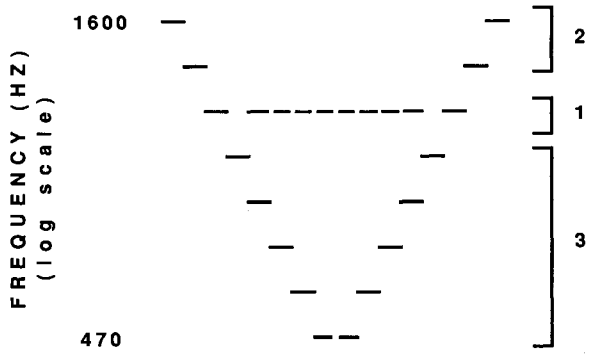


Figure 4.8
Pattern in which there is a competition between grouping based on trajectories and on the proximity of frequencies.

tonal pattern that I made a number of years ago. The pattern is shown in figure 4.8. In it, there are 24 short tones. This pattern was played repeatedly with no gaps between repetitions. Two speeds were used, a faster one of 15.6 tones per second and a slower one of 7.8 tones per second. Three parts of the sequence are separately labeled in the figure although they were not acoustically isolated in any way in the sequence. Part 1 is a set of tones that all have the same frequency and therefore would be expected to be grouped by frequency proximity. Parts 2 and 3 form the upper and lower halves of a pair of trajectories that together form an upright V. When the pattern was played repetitively at the higher speed, parts 1 and 2 formed a single stream on the basis of frequency proximity, leaving part 3 as a separate stream. This partitioning was done on the basis of frequency proximity and this organization became stronger with more repetitions.

When the slower speed was used, the organization became ambiguous. I could hear the frequency-based grouping as in the higher-speed sequence, but I could also hear, if I wished to, the long trajectories formed by perceptually connecting parts 2 and 3. However, I could never hear part 1 alone. The reader may recall that I have postulated that schema-based streaming is asymmetrical. Whereas the subset of sound selected by the schema becomes an perceptually isolated whole, the same does not happen to the "remainder" subset. This was what happened when I organized everything except part 1 by the trajectory principle but could not hear part 1 as a separate sequence. I also postulated that in a competition between primitive grouping and schema-based grouping, speed would favor the primitive grouping. The reader will recall that the frequency-based stream-

ing was stronger with the faster pattern, suggesting that the trajectory-based streaming was based on a schema. I also noticed that the trajectory-based organization grew stronger with repetitions. Since learning will affect the strength of schemas, this effect of repetition points to a schema-governed process.

Experiments have been done that show that sequences of sounds can be made to be more coherent if frequency transitions are introduced so as to connect adjacent sounds. The one that I did with Gary Dannenbring has already been described in chapter 2 and illustrated in figure 2.24.

Some related experiments were performed by Ronald Cole and Brian Scott. They were aware of the fact that unrelated sounds in a rapid sequence tended to segregate into different streams. Why, they asked, did this not happen with speech? After all, consonants such as "s" and vowels are quite different acoustically. Why do they not form separate streams? To answer this question, they noted the fact that as the speech apparatus of a human talker moves from one speech sound to the next, the transition is not instantaneous and gradual changes are seen in the positions of the formants. They proposed that it was these transitions that held the speech stream together. Their first experiment observed stream segregation when a single consonant-vowel syllable was repeated over and over in a loop.[560] An example might be the syllable "sa". A schematic spectrogram of this syllable is shown in figure 4.9, on the left. In one condition, shown on the right, the brief portion of the syllable where the consonant turned into the vowel was cut out and the remaining portions spliced together. They reported that the resulting transitionless syllable sounded quite intelligible when played in isolation. However, when a tape loop was formed in which the syllable was repeated endlessly, after a few cycles the successive repetitions of the consonant (for example, the "s") formed a stream of their own as did the successive occurrences of the vowel. This segregation was, according to the researchers, reported immediately by all listeners. By way of contrast, when they made loops out of the full syllables, leaving the consonant-vowel transition intact, the segregation of consonant from vowel either was not reported at all or was reported only after a much greater number of repetitions.

Concluding that the formant transitions served to hold the stream together, they performed an experiment to determine whether the vowel transitions helped in the task of detecting the order of the basic sounds in a series of syllables.[561] They started with recordings of the syllables "sa", "za", "sha", "va", "ja", "ga", and "fa". Next they separated three components of each sound: the initial consonant, the

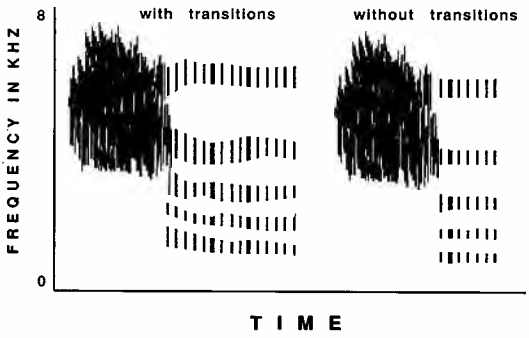


Figure 4.9
Schematic spectrograms of the syllable "sa". 1: with normal transitions. 2: with the formant transitions spliced out. (After Cole and Scott 1973.)

transition, and the vowel steady state. Then they made loops consisting of cycles of four sounds. For example, the formula for one such cycle was [s, sh, v, g]. This meant that the elements in the cycle were extracted from the syllables "sa", "sha", "va", and "ga", respectively. Three types of cycle were made from this formula. The first was composed only of the consonant sections of the four syllables, the second consisted of the four full syllables, and the third consisted of transitionless syllables made from these four syllables. The transitionless and full syllables were equated for duration by adjusting the length of the steady-state portions of the vowels.

The subjects were prefamiliarized with the individual sounds that were used in the loops. Then the loops were played to them and they were asked to write down the order of the sounds in them. They did worst on loops containing only the consonants. Next came the transitionless syllables. The best performance was on the full syllables that contained the transitions, apparently because each syllable was perceived as a coherent unit.

There are a number of ways to interpret these findings. The first is in terms of a trajectory effect. I would direct the reader's attention to figure 4.9. It is tempting to attribute the coherence of the normal syllables to the fact that the formant transitions point backward to positions of high energy in the consonant. To believe that this relation had any organizing value, one would have to suppose that the auditory system could follow a trajectory backward. Retrograde effects are not unthinkable; we have already encountered two examples in the continuity illusion. The first was in the restoration of obliterated words, in which words that come later in the sentence

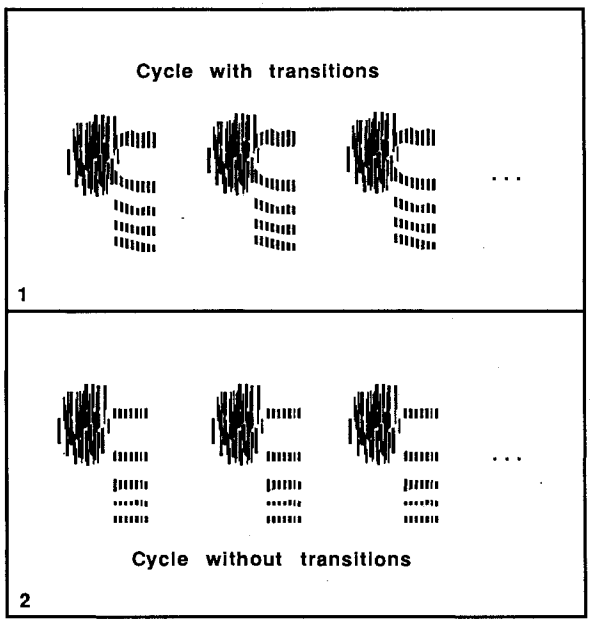


Figure 4.10
Schematic spectrogram of repeating cycles of the type used by Scott and Cole (1973). 1: sequence of normal "sa" syllables. 2: sequence of transitionless syllables.

could determine what the restored sound would be.[562] The second was in the illusory continuity of glides through a noise burst, where the restoration seemed to depend on the relations between the glide that exited from the noise and the glide that entered.[563] Therefore we might entertain the idea that the auditory system can tell that the backward extrapolation of the transitions leads to energy peaks in the consonant, and can conclude that they came from the same source.

A second interpretation of the results does not use the idea of trajectory extrapolation. It says simply that the existence of formant transitions avoids sudden discontinuities in the signal. Continuity is a form of frequency proximity, in which the position of spectral energy in each successive instant is close to where it was at the previous instant. This is an explanation based on continuity, not on the extrapolation of a trajectory.

A third possible explanation says that the elimination of the transitions did not directly affect the grouping of the consonant and vowel within each syllable but favored the grouping of vowels across syllables. To see how this could be so, we should examine figure 4.10,

which shows a schematic drawing of a repeating cycle of "sa" syllables. Panel 1 shows this for a normal syllable and panel 2 for a transitionless syllable. I want to focus not on the role that the transition plays in connecting the vowel to the previous consonant but on its role in isolating successive occurrences of the vowel. When the transitions are present, the formant glides start at a different frequency value than the one that they attain in the steady state. Therefore each successive occurrence of the vowel starts with a spectrum that is different from the final spectrum of its previous occurrence. This difference in spectrum may reduce the vowel-to-vowel grouping that might jump over the intervening consonant. On the other hand, when the transitions are removed, since the vowels are now steady-state, each vowel has the same spectrum at its onset as it had at its previous offset, and this consistency may favor vowel-to-vowel grouping. This is an explanation in terms of the frequency proximity of the nearest temporal parts of successive vowels.

Finally, there is another explanation. The auditory system may be using its knowledge of human articulation to perform the integration. It may convert the acoustical input derived from the sequence of sounds into some representation of the sequence of vocal tract positions that might have produced them. Then, since the sequence is interpretable as the movements of a single speaker's vocal tract, it groups the parts of the syllable together. When the vowel transitions are absent, the hypothesis of a single speaker would be untenable since human vocal tracts cannot snap from one position to another instantaneously. This explanation bases the integration on a schema-driven interpretation of the signal as a speech event.

Although it is not possible to decide in favor of any of these hypotheses from the evidence at hand, the multiplicity of possible explanations makes it impossible for us to accept the data of Cole and Scott as unequivocal support for a primitive trajectory-following process in auditory perception.

In the preceding discussion I have put forward a number of possible alternative explanations that make it hard for us to determine whether the apparent grouping (when it occurs) of sounds that fall on a frequency trajectory is due to a primitive trajectory-following mechanism. Among them were the following:

1. The auditory system may simply detect local discontinuities rather than forming rules that predict the trajectory.
2. Unidirectional ascending or descending sequences may simply be easier for our memories to encode and recall than irregular sequences.

3. It is possible that there is a trajectory-following mechanism but that it is based on schemas and requires prior learning, and, in some cases, a conscious attempt to track the trajectory.
4. There is an absence of competition based on frequency proximity in glissandi.

Because one or all of these explanations may be true, it is hard to verify whether or not primitive auditory scene analysis automatically extrapolates trajectories.

Comparison with Vision The weakness of the trajectory effect in audition contrasts with its strength in vision. This can be seen by examining the phenomenon of crossing trajectories. In this chapter we have seen that if tones alternate between an ascending and a descending trajectory as in figure 4.2 of chapter 2, the listener will never hear the trajectories cross unless the tones on the two trajectories have different timbres. The situation is different in vision. André Achim and I created a visual analog to the trajectory-crossing experiment, substituting the vertical position of a visible spot for auditory frequency.[564] The viewer saw a spot rapidly alternating in position between points on an ascending and a descending trajectory. In other words, figure 4.2 could be used to illustrate the stimuli of the visual experiment if the *y* axis were relabeled "vertical position." As with tones, there was a streaming of the two trajectories, which we called visual stream segregation. The viewer saw an ascending and a descending trajectory of apparent motion happening at the same time, but in contrast with the auditory case, the trajectories crossed one another easily. We did not record whether the actual subjects in our experiment saw the trajectories as crossing or as drawing together and then bouncing apart. However, in our own viewing of the apparatus, we saw both percepts, with the crossing interpretation being at least as common as the bouncing one.

A similar tendency to follow visual trajectories has been reported by Ramachandran and Anstis.[565] The direction of apparent motion of a dot A is ambiguous when two different dots, B and C, are flashed immediately after it. However, they were able to bias the motion toward B rather than C by embedding the AB transition in a longer trajectory of flashed positions leading up to A so that the AB transition continued the direction of motion. They referred to this result as an illustration of "visual momentum" and explained it by the following argument:

> According to Newton's first Law of Motion, a physical object moving at uniform velocity in one direction will persevere in its

> state of uniform motion unless acted upon by an external force to change that state. . . . We have found that any object that moves in one direction will tend to be perceived as continuing its motion in that direction. . . . This might be regarded as a perceptual equivalent of Newton's first law . . . visual momentum may exemplify a prediction by the visual system that at least for small excursions the motion of a physical object is likely to be unidirectional.

This argument sees the existence of visual momentum as an evolutionary adaptation of the visual system to regularities in the physical environment in the manner described so elegantly by Roger Shepard.[566]

Can we formulate a similar argument in audition? I think not. The frequency changes in sounds do not have the same sort of momentum that the positional changes of objects do. One exception to this rule is the Doppler shift in frequency that rapidly moving objects exhibit. However, perceptible Doppler shifts are common only in our technological world with its fire engine sirens, train whistles, and the like, where we often hear very rapidly moving objects, and not in the forests and plains in which the human species took its present form. Also, there is no inertia in a vocal tract that makes it probable that a rising pitch will continue to rise. In fact, the opposite could be true. Since it is likely that the pitch range of the voice is limited, an ascending pitch change is likely to be followed by a descending one. Continuity of change in the voice in the very small scale, though, might be found because it probably takes fewer neural decisions to continue a change than to reverse it; however, the easiest change of all is probably to freeze a frequency transition and continue the sound at a fixed frequency.

Although this argument against pitch-change inertia does not rule out a primitive auditory process for tracking and extrapolating changes, it does not encourage us to expect one. However, human beings can do many things that they are not preprogrammed to do such as ride bicycles, read, and create works of art. These arise from the incredibly complex structure of schemas that must be learned anew by each member of our species. The ability to learn things of incredible complexity makes it possible for the species to adapt to an ever-changing environment. The value of following pitch changes in a world of music, sirens, and the like may well have led to the learning of ascending and descending pitch-change schemas.

"Is Auditory Attention Inherently Rhythmical?"

Another sort of regularity that we should look at is regularity of rhythm. We can ask two questions about it. One is whether there is a tendency for sounds that occur in a regular rhythm to be placed in the same stream by the scene-analysis process. We might imagine that such a tendency would be useful because there are many naturally occurring rhythmic processes in the world. The dripping of a melting icicle, the sound of a walking person, and the rhythmic croaking of frogs all fall into this category. Therefore a repetition of a sound at a constant period might usefully set up an expectation of another repetition of the sound after a similar period.[567]

Jones' theory takes this even further. For her theory, as we saw earlier in this chapter, the process that groups sounds does so by setting up a pattern of rules that predicts the sequence and by controlling the attention of the listener according to those rules. Furthermore, the rules are rhythmical, with the temporal description of inter-event times based on repetitions of a number of basic time periods. These time periods are presumed to be generated by regular oscillatory processes in the nervous system that are used, in combination, for matching the rate of incoming sounds when we are receiving them and for generating the expected rate of sounds when we are anticipating them.[568] For this reason it is claimed by Jones' theory that "attention is inherently rhythmic" and that "temporal predictability may be a prerequisite to the establishment of stream segregation based on frequency relationships."[569] In other words, even a strong segregating factor such as frequency separation would have no effect in segregating the notes of a temporally irregular sequence.

Thus there are three possibilities. One is that rhythmic regularity has no effect on the inclusion of sounds in streams. A second is that although regularity promotes inclusion in the same stream, it is not essential. A final possibility is that temporal regularity is so essential that there can be no stream segregation without it.

This last alternative seems to me to be unlikely, because there are many natural sounds that are not rhythmical that we must nonetheless segregate from other sounds. Examples are the dragging of an object along the ground, the crackling of a fire, and probably even the sound of a voice speaking (although the question of whether speech is or is not rhythmically regular is debatable). However, because Jones' theory takes this position we should consider it as a possibility.

Regularity of Rhythm: Does It Promote Segregation? Let us first consider the question of whether rhythm plays any role in the perceptual grouping of sounds. There is some evidence that it does.

Evidence That Rhythm Favors Segregation The main supporting evidence comes from the work of Jones and her colleagues at Ohio State University. For example, one series of experiments used a pattern of stimuli that resembled those that Alex Rudnicky and I had used earlier to study whether more than one stream can be active at once.[570] The pattern of tones was illustrated in figure 1.7 of chapter 1. In our experiment, the order of two target tones, A and B, was hard to detect in a sequence because the presence of two flanking tones preceding and following the target tones created the sequence FABF, which was hard to discriminate from the alternative sequence FBAF.

The task was made easier when a stream of captor tones (C), near in frequency to the flanker tones, stripped the latter away leaving the target tones in their own stream. In the following diagram we illustrate only the time intervals of the original Bregman-Rudnicky stimulus, not the frequency separations. The frequencies had a pattern similar to the one shown in figure 1.7.

C C C F A B F C C C C

However, Jones and her co-workers pointed out that not only were the captor tones close in frequency to the flanking tones that they were to capture, but they also were rhythmically related to the flankers such that the captors and the flankers considered together formed a slower isochronous sequence than the one formed by A and B. This can be seen in the preceding diagram. They argued that A and B might have been segregated by this difference in rhythm. Therefore they performed a number of manipulations of the time intervals between the tones so as to create other rhythms. For example, in one condition that can be roughly illustrated by the following diagram, the F's fell into an isochronous sequence with AB but not with the C's.

C C C F A B F C C C

In another sequence the entire pattern formed a single isochronous sequence:

C C C F A B F C C C

In both of these cases, the AB sequence was harder to isolate than it was in the original pattern. The researchers concluded that "temporal predictability may be a prerequisite to the establishment of stream segregation based on frequency relationships."[571] I interpret them as proposing that frequency differences would cause streams to segregate only if each stream could follow a regular, and preferably different, rhythm. Although these results did not really warrant such

strong conclusions they did demonstrate some role of temporal patterning in the ability to attend to only some of the tones in a sequence.

Perhaps, rather than governing the process of primitive grouping, the rhythmic relations affected the ability of the listeners to ready their attention for the critical pair of tones whose order was to be judged. We should note that in Bregman and Rudnicky's experiment there were two influences at work. The first was the voluntary attention of the listener which was trying to select out the target tones, and the second was the involuntary primitive grouping of the tones, which was opposing the efforts of attention. It is this opposition that tells us that primitive grouping exists independently of attention. Rhythmic isolation may assist the selection process *directly* rather than indirectly through its effects on primitive grouping.

This conclusion is supported by the results of another study by Jones and her colleagues.[572] The listeners heard pairs of nine-tone auditory sequences, the second one transposed up or down as a whole relative to the first. In addition to the transposition, one of the tones in the second sequence was altered in pitch on half the trials. The listeners were required to judge whether the sequence was merely transposed or altered as well. At 3.3 tones per second, the sequence was slow enough that there would have been no compelling stream segregation. Yet the subjects could more easily detect the change of a note when their attention was guided there by the fact that the rhythm of the sequence placed an accent on this note. In this experiment it was probably the encoding or mental description of a sequence, rather than its primitive segregation, that was affected by the rhythm.

A similar conclusion about the effects of rhythmic grouping can be reached from the results of an experiment by Dowling.[573] His subjects heard a sequence composed of four phrases each consisting of five tones. They can be diagrammed as follows:

ABCDE FGHIJ KLMNO PQRST

The separation between the phrases was achieved by lengthening the last note in each one and adding a short silence (if the first four notes are thought of as quarter notes, the fifth can be thought of as a dotted quarter note which was followed by an eighth rest). Immediately afterward, the subject was presented with a test pattern of five tones and was asked whether it had occurred in the original sequence. The tests consisted of two types, one in which the tones consisted of a whole phrase from the original sequence (for example, FGHIJ) and the other in which the tones bridged phrase boundaries (for example,

IJ KLM). The test sequences that preserved the original phrasing were much easier to recognize. This result shows that an essential part of the memory for a sequence is its rhythmic pattern.

The question of whether rhythm affects the formation of streams was taken up by Stephen Handel and his fellow researchers at the University of Tennessee.[574] They manipulated the rhythm of a four-tone sequence, presented as a repeating cycle, by varying the positions of silences in it. If we represent the frequencies of four tones of increasing frequency by the digits 1 to 4 and silences by hyphens, the formula 12–34– would represent a sequence in which the two low tones and the two high tones were placed in separate groups, separated by silences. The formula 1–234– would isolate the lowest tone into a separate group. The experiment also varied the frequency separation between the tones. Each cycle of four tones (together with whatever silences were added) took 0.8 second, yielding an average tone rate of 5 per second.

This experiment is not really relevant to the question of whether predictability per se influences the formation of streams, because when a short sequence is recycled over and over, it rapidly becomes very predictable. In addition, any short repeating sequence can be described by a few rules and is therefore regular in Jones's sense. The study looked at the effects of particular forms of temporal grouping on the nature of the resulting streams.

The first finding of note was that the temporal grouping had little or no effect when the sequence consisted of two high and two low tones, where the high pair was separated from the low pair by a considerable frequency difference (400 and 500 versus 1,600 and 2,000 Hz). Apparently the frequency separation was the dominant influence on streaming.

When the four tones were equally spaced by octaves (750, 1,500, 3,000, and 6,000 Hz), the temporal grouping did matter. When the sequence was isochronous, the four tones tended to be heard as a single unsegregated sequence. When silences were added, tones preferred to group with others that were closest to them both in frequency and in time. Thus, rhythmic differences had an effect because they were really differences in temporal separation.

There was one exception to this general rule. There was a greater tendency to break up the four-tone sequence into separate perceptual streams when the resulting streams would be isochronous. If this should be borne out in future studies it means that a stream is stronger when its members are equally spaced in time and confirms Jones's hypothesis that rhythmic processes can be involved in the formation of streams. Notice, however, that this factor was evident only when

large frequency separations did not force the streaming. Therefore these results do not lend support to the proposal by Jones and her associates that substreams will form *only* when they are rhythmically regular.

Segregation Occurs with Temporally Irregular Sequences Although there is some evidence that rhythmic regularity assists the formation of streams, there is enough evidence to reject the hypothesis that it is essential for their formation. It is not unusual to find, for example, a case in which an isochronous sequence breaks into two streams on the basis of frequency proximities, and neither of the two streams formed in this way is isochronous.[575] Another experiment that I reviewed earlier, by Tougas and Bregman, also showed that rhythmic regularity is not essential.[576] When a falling sequence of tones crossed a rising sequence, the listeners tended not to be able to follow either sequence across the crossing point, but switched over to the part of the other sequence that fell in the same frequency region as the earlier part of the one they had been following. There were two rhythmic conditions: In one, the sequence would remain isochronous if the listeners switched their attention to the other trajectory and in the other it would be isochronous if they continued to follow the same trajectory. The tendency to switch trajectories was unaffected by this rhythmic factor and seemed entirely governed by the principle of frequency proximity.

A recent experiment performed by Marilyn French-St. George and myself looked specifically for whether predictability in the pattern of pitches and time intervals would help a listener to integrate a sequence that was being segregated by frequency differences.[577] It found no such effect. The listeners were asked to listen to a 30-second sequence of pure tones, each of which was 40 msec in duration, with an average rate of about 11 per second. There were four main conditions.

In condition A the sequence was very regular. It was formed of a repeating cycle of eight tones, four lower ones and four higher ones. The onset-to-onset time between successive tones was always exactly 91.2 msec. The four higher tones were spaced apart by semitone separations, thus spanning a three-semitone interval, and the lower tones were spaced from one another in the same way. The high tones were interleaved with the low tones in the cycle. The listeners were told that during the 30-second presentation of the cycle they were to try to hold the tones together as a coherent sequence and to indicate when they were successful by pressing a key in front of them. On different trials, each lasting 30 seconds, the separation between the mean frequency of the higher and the lower tones was varied from

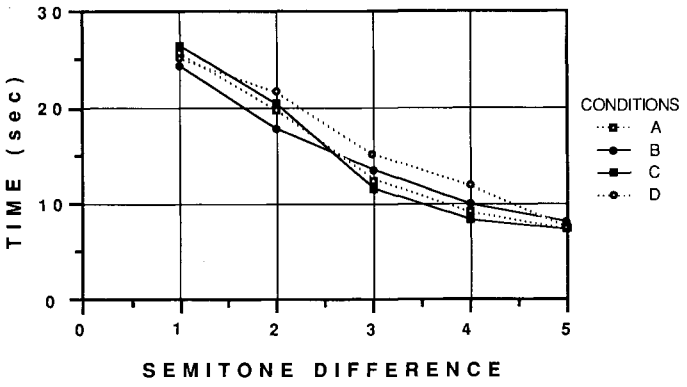


Figure 4.11
Results of an experiment by French-St. George and Bregman (1989) on stream segregation in which frequency separation and predictability were varied.

one semitone to five. In the one-semitone separation the two frequency ranges overlapped considerably, while for the four- and five-semitone separations they did not overlap at all.

In condition B the listener could predict when the next tone would come, but not what its frequency would be. In this condition, the same set of high and low tones was used as in the regular condition but the 30-second sequence was not a repeating cycle. The order of each successive eight tones was computed independently by shuffling the order of the high tones, shuffling the order of the low tones, and then interleaving them, so as to maintain the high-low alternation of the sequence. Again, the frequency separation between higher and lower tone sets was varied across trials.

In condition C the frequency of each successive tone was predictable, but the time between the tones was irregular. The sequence was a fixed repeating cycle of the original eight frequencies but the onset-to-onset time between successive tones was randomized between limits of 65.6 msec and 116.8 msec so as to keep an average rate of 91.2 msec. Again each 30-second trial had a different frequency separation.

In condition D we randomized the sequence of frequencies as in the B condition and the sequence of time intervals as in C.

In looking at the results, we found that that, in all conditions, the longer the listener listened, the less coherent the sequence became. This is in agreement with previous findings.[578] The results concerning frequency separation and predictability are shown in figure 4.11. In all four conditions we found, as expected, that the mean number of

seconds, out of 30, for which the listener could hear the sequence as a single stream became less as the frequency separation between higher and lower tones became greater. Although Jones, Kidd, and Wetzel have proposed that stream integration is a prediction-based process and that the effect of frequency separation is restricted to cases where the sequence is predictable, we found no such restriction in our data.[579] Furthermore, the predictability of the sequence did not seem to help at all in holding the sequence together. The effects of frequency seemed about the same in all the conditions. If there was any condition that was different from the others it was D, the one in which both frequency and time were unpredictable; yet this sequence was easier, not harder, to hold together. The superiority of this condition, however, was not statistically significant.

It is still possible that predictability can help hold a stream together and that this experiment did not detect the conditions under which it does. For example, whenever the total stream was predictable, the substreams (high and low) also were. We do not know whether the predictability of the sub-streams helped them be heard as separate units, thereby counteracting the effects of the regularity of the global sequence (even though the listeners were asked to hold the overall sequence together). We need further experiments in which the regularity of the overall sequence and that of the within-substream sequence is varied independently. Another possible criticism of this experiment might be made: Due to the fact that the frequency separation changed on each trial, the listeners did not really become familiar enough with the regular sequence to take advantage of its regularity. In conditions with large frequency separations the sequence might have split before the hearer had a chance to learn it. Yet in certain intermediate separations, the listeners were able to hold on to the sequence as a coherent stream for a while, though not for the full 30 seconds. This should have enabled them to learn the regular sequence, yet there is no advantage for the regular conditions even at these intermediate separations. Another possibility is that the range of uncertainty in frequency and timing was too small.

Despite these criticisms, the experiment did succeed in showing that, at least within a certain range, neither a predictability in frequency nor in rhythm is a requirement for obtaining stream segregation effects.

The preceding sections have asked whether the regularity of a sequence, whether of pitch sequence or of timing, is responsible for holding it together. A number of observations have shown that while regularity seems in some cases to assist the attention in tracking a sequence of sounds over time, regularity is by no means a

requirement for something to be heard as a coherent stream. Indeed, if it were, we would never be able to hear irregular sequences of sounds and we would be very surprised to discover that such sequences existed.

We have considered two elementary forms of regularity, simple trajectories and simple isochronous rhythms. Neither has a dominating influence on the formation of streams. Furthermore, it is not clear how to interpret the influences that do exist. Do they affect the primitive grouping process itself or some higher-level processes such as memory or attention?

"Are Streams Created by Attention?" Of course one could take the approach of denying that such a division existed and assert, instead, that all those phenomena that I have interpreted as arising from the activity of a preattentive process of primitive grouping arise instead from the activity of the attentional system itself. This approach would argue that a stream is simply that sequence of sounds that is selected by attention and that all the factors that are thought to influence primitive grouping are really influencing attention. This would imply that there is only one stream active at any moment, the one that we are attending to. However, the Bregman-Rudnicky experiment was done to examine this question and it concluded that streams were being formed among tones that were being rejected by attention and that this made it easier to reject them as a group (the "wrap up all your garbage in the same bundle" heuristic). I would refer the reader to the many discussions of this experiment elsewhere in this volume, especially the one in chapter 2, where we consider the question of whether there is only one stream active at a time.

Helmholtz might well have subscribed to the view that grouping is due to attention. For example, when he presented his method of playing a tone before a complex tone to assist a listener in hearing out one of the partials, he described it as a way of directing the listener's attention to the relevant partial. I have presented this method, on the other hand, as being derivable from a very general principle of preattentive grouping.

Could either Helmholtz or the author be wrong? To avoid such an outcome let me restate an earlier explanation that, while inelegant, has the virtue of preserving the honor of two scientists. There are both preattentive and attentional processes that can latch onto acoustic features as a method of attaining their ends. In the example of hearing out a partial, this process seems to involve effort and trying to hear the earlier sound in the mixture. This suggests a role for conscious attention. On the other hand, acoustic factors can make this

attempt easy or hard. While it is possible that they affect the difficulty by denying the attention a clear criterion for latching onto the desired sound in the mixture, I believe that the existence of a primitive scene-analyzing process (supported by many arguments in this volume) is an attractive enough idea to be worth keeping.

There is merit in having redundant systems, a primitive one and another that is schema-driven. The primitive one can make early learning possible in a wide variety of environments and the educated one can sharpen up our listening in environments whose particular regularities we have learned to exploit. Let us think of a schema as a pattern of regularity in the world as it has been captured by some formula in the brain.

It might be instructive to ask ourselves what kinds of schemas are used by a listener who is trying to use an earlier sound to pull out a partial from a mixture. There are obviously all the myriad schemas by which we control our behaviors, including ones for decoding the language of the scientist who has sat us down to do this unlikely task. However, a crucial schema for doing this particular task is the notion of a "repetition." Without it we could not program ourselves to listen for something in the mixture that was a repetition of the earlier sound. I am not saying that we could not *hear* a repetition without a cognitive representation of the concept of repetition. I am saying, however, that we could not *listen* for one. That is, we could not govern our attention in a goal-directed way. The topic of schemas is far too rich to be dealt with in this book. I mean only to give a flavor here of what we would be looking for if we began to look for schemas that might help us to do auditory tasks.

I do not mean to imply that the two proposed processes, the one primitive and preattentive and the other schema-driven and attention-directing, always react the same way to the same acoustic variables. If this were true there would be no way to distinguish them. Sometimes they react differently. For example, when you are trying to hold on to a one-stream interpretation of a sequence of alternating high and low tones, the frequency separation of the tones and the speed of the sequence interact. At faster speeds a smaller frequency separation is required, presumably because a primitive process is trying to form substreams at faster speeds while the attentional process is trying to hold the sequence together. However, if you are trying to segregate the two streams, your ability to do so requires only a small frequency separation and is virtually independent of the speed of the sequence. Apparently when the primitive process wants to take the sequence apart, this hurts the voluntary segregation process. But even when the primitive process creates only a single

FAMILIARITY BREEDS CONTEMPT

FAMILIARITY BREEDS CONFTLEEMAPST

Figure 4.12
Extracting a set of letters out of a mixture by using the fact that they complete a familiar phrase.

stream, as when the frequencies are close together, it is still possible for the attentional process to take it apart by listening in a goal-directed way for either the higher or lower tone. These observations seem to show that it is easier to *select* a part of a sequence against the dictates of the primitive process than to *integrate* it against the workings of the primitive process. Attention does not operate with equal facility on any combination of sensory elements. Otherwise it would not be hard to select anything at all for attention. Hopefully we can use these conflicts between what we want to do and what we can do to clarify the nature of primitive grouping and attention.

I have taken the position that primitive grouping processes actually sort the signal but that the more sophisticated schema-based processes simply select what they need and do not leave a residual that is either diminished in any way or organized. In some sense they do not remove the evidence that they need from a mixture. They interpret it, but leave it there for other schemas to interpret too. It would be possible to design an experiment to test this view. It would be necessary to do an experiment along the lines of the one done by Bregman and Rudnicky that I described in chapter 1 and illustrated in figure 1.7.[580] In that experiment, a sequence of tones formed a stream and captured target tones out of a second stream in such a way as to make a judgment task on the second stream easier. This latter effect proved that it was a true capturing that leaves a residual that is simpler than it would be without the capturing. Now the experiment must be altered so that the target tones are not bound to their captors by frequency similarity, as they were in Bregman and Rudnicky's experiment, but by the fact that captors and targets together form a familiar pattern, one that would plausibly be united by means of a schema. It would have to be similar to the sequence that I have illustrated in figure 4.12, with sounds taking the place of letters.

Although I made up the example myself and the second line is the same as the first, I can extract the second word "contempt" from the jumble of letters only with effort, and I cannot detect the residual

COMPUTATIONAL

Figure 4.13
Segregation of letters by the primitive properties of size and darkness.

word without an elaborate mental process of calculation. Look at the difference when the segregation is induced by the primitive factors of size and darkness, as shown in figure 4.13. The segregation is now effortless and both words in the mixed sequence are legible. I would expect the auditory experiment to show a similar effect.

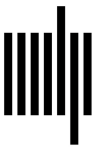
Auditory Scene Analysis

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