Chapter 3

Integration of Simultaneous Auditory Components

The task of segregating sounds that had arisen from different sources was described in chapter 1 through the example of the spectrogram of a mixture of sounds (see figure 1.4). It was apparent that there are two kinds of grouping to be done. First, a process of sequential grouping was needed to put some of the spectral components that followed one another in time into the same perceptual stream, where they could be used to calculate the sequential properties of a source of sound in the environment (such as its melody or rhythm). This process was the subject of chapter 2.

A second process was needed to partition the set of concurrent components into distinct subsets, and to place them into different streams where they could be used to calculate the spectral properties of distinct sources of sound (such as timbre or pitch). It is this process that will be the subject of the present chapter.

The question of how to study the process of simultaneous grouping and segregation is not obvious. However, in the summer of 1975, Alex Rudnicky and I began work on a set of auditory demonstrations that were related to the question of whether two tones could be overlapped in time and still separate into two perceptual components, as opposed to becoming a simple complex tone. The first demonstration made use of a repeating cycle of two discrete tones, A and B, each of the same fixed duration and a fixed rate of repetition. The two tones were at different frequencies, but overlapped in time as shown in figure 3.1. The percentage of the length of the tone that overlapped the other tone was varied. We found that if we used tones whose duration was, for example, 250 msec per tone, and that were only 25 percent or 50 percent overlapped (i.e., relatively asynchronous in onset) the tones would segregate clearly into separate streams. When the temporal overlap reached 88 percent (almost synchronous onsets) the tones fused into one percept. When the two tones fused the resultant tone sounded complex, but when they segregated each sounded pure. Segregation was clearer when the frequency separation was

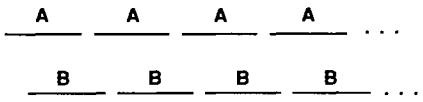


Figure 3.1
Repeating cycle of overlapping tones.

greater; this suggested to us that some sort of stream segregation phenomenon was involved.

This demonstration used tones with silences between them. We decided to see whether two simultaneous and continuous tones of different frequencies could be induced to segregate or fuse depending on brief amplitude changes that were synchronous or asynchronous for the two frequencies. We introduced brief amplitude drops (8 per second) in the two, otherwise continuous, tones, causing a warbling sound in each tone. This is illustrated in figure 3.2, with the two dashed lines representing the tones and the V's representing the dips in loudness. (In the example the dips are not synchronous.) We could not cause the tones to totally segregate on the basis of asynchronous warbling. When the warbles were at different rates for the two frequencies, the two different rates could be heard, but the tones were not really perceptually segregated; the quality of each one was affected by the presence of the other.

So far, we had seen that when each component switched completely on and off, with a sizable off period, stream separation was high under asynchronous conditions; however, where the amplitude merely dropped off briefly, asynchrony did not produce complete separation. We then decided to confine our attention to an intermediate condition, which used 224-msec tones with 32-msec silences between repetitions. When the tones had asynchronous onsets in this condition, we were able to perceptually pick out the top or bottom stream. An analogy to our earlier studies of stream segregation was noted: *The relatively pure-sounding tones of the streams emerged only after several cycles, a period of time that apparently was required for the buildup of the stream criteria.*

We next made a series of observations employing two tones, a high and a low, repeating at different rates, e.g., 5 tones per second for the high tone and 4 tones per second for the low tone. Thus, once per second the high and low tones would have synchronous onsets; otherwise they were out of synchrony. The perceptual experience consisted of two streams, but with some interesting effects. Every time the onset synchrony occurred, a rich tone was heard in the lower

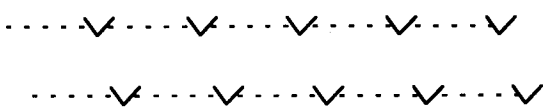


Figure 3.2
Partial segregation with brief amplitude drops (represented by V's).

stream and the corresponding tone in the upper stream was weak or missing. The lower tone was heard to be louder as well. The lower tone percept thus seemed to capture the energy of the synchronous higher tone. This reminded us of the exclusive allocation principle in perception which states that a piece of perceptual evidence has to be allocated to one perceptual entity or another, but not to more than one at a time. Here a certain amount of the energy was being subtracted from the upper tone and added to the lower one.

We then went on to listen to cycles that contained repetitions of three tones of different frequencies that might or might not overlap in time. In this example, it was possible to hear three streams. Asynchronous onsets of three tones did produce segregation, but not as strongly as with only two tones. Nonetheless we made an interesting observation. One of the patterns in this series included a temporal point at which a tone in all three streams had a simultaneous onset. At this point a strong "vertical" fusion took place. This fusion was different from the one we had encountered in the two-stream situation in that now the fusion was so compelling that it rendered the new fused unit quite perceptually independent of the single-tone streams, and did not simply color the timbre of the tone in the lowest stream.

We should note that since only about one-eighth of each stream was composed of silence, most of the time all three frequencies were present simultaneously. Hence all combination tones, beat phenomena, dissonance, and so on, were occurring most of the time. Yet when the streams segregated they sounded pure, all those acoustic combination effects being ignored by the ear. Apparently when the contributing tones were assigned to separate perceptual objects (streams) the awareness of these interactions was suppressed. We might ask, if this is the case, why the ear ever hears these interaction phenomena if it is capable of rejecting them. The answer may depend upon the fact that in some cases, such as when the combination effects are caused by the interaction of acoustic components that all belong to the same real-world sound, these phenomena are as characteristic of that source of sound as the pure-tone components are and therefore should be incorporated into the description.

But how can our auditory systems tell the difference between acoustic phenomena that are an intrinsic part of a sound and those that arise from the accidental co-occurrence of two unrelated sounds? It must obviously fall back on some of the same techniques that it uses to decide whether pure-tone components belong to the same sound. These cues may act to bind the interaction effects into one or another stream. Take, for example, a case in which two complex tones, each composed of a simple harmonic series and therefore heard as having no internal dissonance, are sounded at the same moment. In this case the combination phenomena have a synchronous onset with the pure-tone components, and therefore are also captured into the stream, adding to its timbre or perceptual complexity. On the other hand, if one of these tones occurs for a short time by itself and then is joined by a second tone, the interaction properties will not appear until the second tone is sounded and therefore the onset of the interaction phenomena starts in synchrony with the second tone but not the first. Under these circumstances, the richness or dissonance would be expected to be assigned to the second tone but not the first. However, if the second tone lasted longer than the first, so that it appeared at the end without any interaction components, the auditory system would have evidence that it, too, was pure and might not assign the interaction properties to it either.

The most important observations of Rudnicky and myself could be summarized by saying that the onset synchrony of elements in two streams increased the perceived intensity of the element in the lower stream and decreased the intensity of the one in the upper stream, and that this effect decreased with an increase of frequency separation between the two simultaneous tones.

Although the experimenters were trained listeners and could describe the phenomena reliably to one another, the requirement of replicating the observations with naive subjects required us to reduce the task to one involving simple comparisons or descriptions. For this reason, the experiment by Bregman and Pinker was undertaken.[284]

A Miniature Scene-Analysis Problem

This experiment has already been described in chapter 1 and illustrated in figure 1.16. It involved the sequential capturing of one component from within a two-component tone. It used a cycle in which a pure tone A was followed by a pair of more-or-less simultaneous pure tones B and C. The latter pair could be viewed as forming a complex tone BC. The timing was as follows: 117 msec of silence, A for 117 msec, 47 msec of silence, and then BC, with each of tones B

and C lasting for 117 msec. This pattern was repeated cyclically. In figure 1.16, tone A is just a little higher in frequency than B and tone C comes on a bit earlier than B; however, not all the conditions were like this.

The cycle could be heard in more than one way depending on the frequencies and timing of the three-pure tone components. One way was as an alternation of a pure tone A and a rich tone BC, the repetitions of each of these belonging to a stream of its own. Another way was to hear the pure tones A and B in one stream (AB AB AB. . .) and the isolated pure tone C in another (C C C. . .). There were, of course, some ambiguous cases in which the listener experienced some mixture of these two percepts. The way the listener heard the sounds depended on two factors that were manipulated in the experiment: (1) the frequency separation between tones A and B, tone A varying in frequency from a semitone above B to 1.7 octaves above, and (2) the asynchrony of onset of B and C, tone C either coming on synchronously with B or else leading it or lagging it by 29 or 58 msec (25 or 50 percent of the length of one tone).

In one experiment the listeners were asked to judge the richness of the timbre of tone C. "Timbre" was defined for them as "the quality that distinguishes, for example, the sounds of different musical instruments producing notes of the same pitch." They were asked to judge this richness in relation to two standard tones that were presented before each judgment was made. The "pure" standard tone contained only tone C, but the "rich" one contained both B and C. In the rich standard tone, the richness of the BC complex was heard because there was no capturing tone to break it up. In effect the subjects were being asked how much the C tone sounded like it did when embedded in a BC complex, versus sounding like it did when played in isolation. It was assumed that if B fused with C this would make C sound richer.

When tone C was less synchronous in onset (and therefore in offset) with tone B, the former was heard as more pure. We interpreted this as meaning that C fused less with B when the two were not synchronous. However, a critic might have offered a simpler explanation. Perhaps when C was not synchronous with B, the listener could hear C alone for a brief moment and this was sufficient to reduce the richness ratings that were given to it. However, there was also a more remarkable result that could not be explained in this way: when A was closer in frequency to B, tone C sounded purer. We explained this by arguing that when A was close to B in frequency, it captured B into a sequential stream (AB AB. . .). This tended to remove B

from its perceptual association with C, and this, in turn, released C to be heard as a pure tone.

In a separate experiment, the listeners were asked to make a different judgment: whether A and B were heard in the same stream. Of course, in order to be able to do this the listeners had to be trained on what it meant to hear one stream as opposed to two. Their training consisted of a tonal sequence that contained a rapid alternation of a higher and a lower tone. At first the tones were near one another in frequency and sounded like a single coherent sequence, but gradually they moved apart in frequency until they split perceptually into two streams.

The main experiment yielded two results. The first was that when the frequencies of A and B were more different, the two tones sounded less like they were in the same stream. We interpreted this as showing that the capturing of B depended on the frequency proximity between A and B. However, as in the previous experiment, a simpler explanation might have been sufficient. Since the listeners were trained to call a stream "single" when its members were close together in frequency, they might have simply used this criterion, rather than the actual stream segregation, to make their judgments. Fortunately, just as in the previous experiment, there was a second result that called for an explanation in terms of perceptual grouping: the synchrony of B and C affected the perception of A. In conditions in which B and C were less synchronous in onset and offset, A and B were more strongly judged as being in the same stream. Our explanation for this result was that when B and C were asynchronous, they tended not to fuse as well into a single sound. This released B so that it could more easily be captured by A into a sequential AB stream.

The results of these experiments seem to be describable in terms of the competition between two types of organization. One is a sequential process that groups A and B into the same stream, with the strength of their grouping depending on how similar they are. The second is one that fuses B and C into the same simultaneous grouping and does so with a strength that depends on how synchronous they are. The effect of the competition is such that it takes less strength to capture a component into a sequential organization if it is less strongly bound into a simultaneous organization (and conversely).

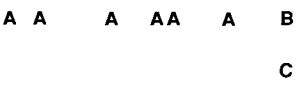


Figure 3.3
Effect of a sequence of captors.

My interpretation of the Bregman-Pinker experiment using the ABC pattern depends on the assumption that it is the same sequential grouping process at work in their experiment as in the experiments on sequential streaming that were described in chapter 2. If this assumption is true then several predictions can be made for experiments using the ABC pattern.

1. Increasing the rate of the sequence (or perhaps decreasing the distance between the offset of A and the onset of B) should strengthen the AB grouping. The former effect has been observed by van Noorden.[285]
2. Because sequential grouping builds up with repetition of components in a given frequency region, increasing the length of time that one listens to the cycle should increase the AB grouping.
3. For the same reason as given in the previous prediction, if a single occurrence of Bregman and Pinker's ABC pattern were preceded by a long string of A's as shown in figure 3.3, then as this string was made longer, the AB grouping should become stronger (and the BC grouping thereby weakened). (In the figure the timing of the captors is irregular to avoid effects of rhythmic regularity.)
4. The next prediction applies to the pattern shown in figure 3.4. If a three-component sound (call it XYZ) were repeatedly alternated with a simpler tone, XY, where two of the partials in XYZ appear in XY, XYZ should be divided perceptually into two pieces, XY and Z. (This might be more likely if the two parts of XY were more integrated, e.g., by starting synchronously.)

Although these predictions are in agreement with informal observations that I have made, as far as I am aware there has been no formal experimentation that confirms them (except where noted).

Before we leave our discussion of the ABC experiment, we should consider a possible criticism. How could we answer a critic who

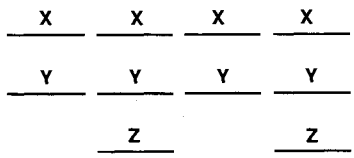


Figure 3.4
Decomposition of a three-component tone through capturing.

claimed that when we heard tone A repeating again as part of B, it was just a figment of our imaginations, a memory of A that was somehow being activated again by the BC mixture? Helmholtz raised the question and answered it in 1859.[286] He was discussing the fact that he could help himself hear a harmonic of a complex piano tone by preceding it with a tone whose fundamental was at the frequency of the desired harmonic. He pointed out that when the prior tone was very close but not identical in frequency to the desired harmonic, he could hear that the prior tone and the harmonic were not the same. Furthermore, if the complex tone did not contain a component that was close to the priming tone in frequency, no harmonic would be heard. A result similar to the latter was reported by van Noorden when he alternated a pure and a complex tone rapidly.[287] We can see that a frequency identical to the captor's is heard in the complex tone only when it is really there.

Another demonstration that it could not be just a ringing in the ears comes from a repetition of the ABC experiment using as A, B, and C, tones that glide from one frequency to another rather than tones that remain steady at one frequency. Howard Steiger and I carried out this experiment.[288] The stimulus was shown earlier in figure 2.17 of chapter 2. A pure tone glide X alternated with a complex glide whose components were Y and Z. Glides Y and Z moved in parallel on a log frequency scale. X acted as a captor that, under certain conditions, could capture Y into a sequential stream so that Y no longer fused with Z. This capturing was strongest when Y was an exact copy of X. Then the listener heard X occurring again in the mixture. Clearly, since Y was a complex event, hearing it could not be explained as "hearing a ringing in ones ears." If by a ringing, we mean a replaying of the memory of X, that memory must be complex enough to encode a glide.

Even if we are not just replaying our memory of X but actually pulling Y out of the mixture YZ, we know that the memory of X that is influencing this decomposition must involve the glide property. We know this because the most effective captor of a glide is another identical glide. A steady-frequency tone, matching either Y's mean log frequency or the frequency of its beginning or end points, is not as effective. If the only part of the memory of X that counted was some single fact about its frequency, a steady-frequency X should serve just as well. The effectiveness of the capturing, however, even depended on the match in the direction of gliding of X and Y. If X was ascending and Y was descending, both centered on the same frequency, the capturing of Y was weaker than if both were ascending

or descending. We can conclude that there is a fairly complex memory for X that is employed to look for near matches to X at subsequent moments of time.

It is important at this point to recall the significance of the ABC experiment as a model of the scene analysis process that we have been discussing. In ordinary listening we have to group acoustic components both sequentially and simultaneously in order to allocate them appropriately to streams that represent individual sources of sound. The ABC experiment provides a glimpse of this process working on a very simplified problem. Because of the simplification, we cannot say that the factors that controlled grouping were representative of those that work in real life. However, we have learned that the factors of frequency proximity and synchrony are among those that are usable and have also learned about the competition between the two types of grouping.

Factors Influencing Integration of Simultaneous Components

The discussion of the process of the integration of simultaneously present acoustic components divides naturally into two questions. The first of these asks what the acoustic clues are that tell the auditory system which components have come from the same physical event. The second question asks about the perceptual consequences of having grouped spectral components; what features of our perceptual experience depend on this grouping? The first question is the topic of the following sections.

How do we know which acoustic components have arisen simultaneously from the same physical event? The answer that I would like to propose is this: If a group of components have arisen from the same physical event, they will have relationships between them that are unlikely to have occurred by chance. The relationships that I am thinking of are of an entirely physical nature and concern properties such as timing, frequency, and the differential effects on our two ears. They are the consequence of the physical laws that govern the sounds and their effects on our ears. We can view the scene-analysis system as trying to take advantage of these physical relationships in order to put the components together appropriately. In effect, they are trying to run the laws of acoustics backwards to reconstruct the original sounds. It is unlikely, however, that the grouping principles directly resemble the physical laws that have created the sounds. Rather, the principles have evolved to take advantage of relationships that the laws create. This section, then, will discuss those relationships that

are likely to exist between simultaneous components that have arisen from the same physical event, but are unlikely to exist between those that have arisen from unrelated events.

The "Old-Plus-New" Heuristic

The results of the Bregman-Pinker experiment suggest that the auditory system uses a rule that we might label "the old-plus-new heuristic." It could be stated as follows: "If you can plausibly interpret any part of a current group of acoustic components as a continuation of a sound that just occurred, do so and remove it from the mixture. Then take the difference between the current sound and the previous sound as the new group to be analyzed." I use the word heuristic in the same sense as it is used in the field of artificial intelligence, as a procedure that tends to give the right answers, or at least improve the quality of the answers, but is not guaranteed to do so in all cases.

As we find with many principles of audition, when we think we are the first to notice something we discover that Helmholtz, the great nineteenth-century German physicist, was ahead of us. So it is for the old-plus-new principle. Helmholtz was a strong supporter of the idea, proposed by the physicist G. S. Ohm, that the auditory system registers the individual spectral components of a sound. Realizing the implications of this idea, Helmholtz knew that in the case of acoustic mixtures, there would be a problem in deciding which components had arisen from the same source. Here is an excerpt from a passage that describes a person listening to a group of musical instruments. It describes the old-plus-new heuristic clearly.

> Now there are many circumstances which assist us first in separating the musical tones arising from different sources, and secondly, in keeping together the partial tones of each separate source. Thus when one musical tone is heard for some time before being joined by the second, and when the second continues after the first has ceased, the separation in sound is facilitated by the succession of time. We have already heard the first musical tone by itself, and hence know immediately what we have to deduct from the compound effect for the effect of this first tone. Even when several parts proceed in the same rhythm in polyphonic music, the mode in which the tones of different instruments and voices commence, the nature of their increase in force, the certainty with which they are held, and the manner in which they die off, are generally slightly different for each. . . . When, then such instruments are sounded together there are generally

> points of time when one or the other is predominant, and it is consequently easily distinguished by the ear.

In the same passage Helmholtz describes other factors that promote appropriate separation of sounds. They will all turn out to be important in the discussion that follows.

> But besides all this, in good part music, especial care is taken to facilitate the separation of the parts by the ear. In polyphonic music proper, where each part has its own distinct melody, a principle means of clearly separating the progression of each part has always consisted in making them proceed in different rhythms and on different divisions of the bars; or where this could not be done . . . as in four-part chorales, it is an old rule, contrived for this purpose, to let three parts, if possible, move by single degrees of the scale, and let the fourth leap over several.

Helmholtz also goes on to point out that the factors that distinguish the partials of different sounds that have been mixed together will not segregate the partials of a single sound.

> All these helps fail in the resolution of musical tones into their constituent partials. When a compound tone commences to sound, all its partial tones commence with the same comparative strength; when it swells, all of them generally swell uniformly; when it ceases, all cease simultaneously. Hence no opportunity is generally given for hearing them separately and independently.[289]

Helmholtz, however, gives the reader a special method for hearing out the third harmonic of a *c* played on the piano:

> In commencing to observe upper partial tones, it is advisable just before producing the musical tone itself which you wish to analyze, to sound the tone you wish to distinguish in it, very gently, and if possible in the same quality of tone as the compound itself. . . . First gently strike on a piano the note g' [the g in the octave above the c that is to be analyzed] . . . and after letting the [piano key] rise so as to damp the string, strike the note c, of which g' is the third partial, with great force, and keep your attention directed to the pitch of the g' which you had just heard, and you will hear it again in the compound tone of c.[290]

He showed how by the use of a similar method a listener could hear individual partials in different sorts of instruments and even in the human voice. Notice how closely Helmholtz's method resembles the

method used by Bregman and Pinker. Helmholtz was not proposing a full-blown theory of perception and would probably have been amazed at our giving a name such as the old-plus-new heuristic to what seemed like a commonsense method for directing the listener's attention to the appropriate part of the spectrum. In his view, no doubt, he was simply using his knowledge of the physical structure of sound and his experience in the laboratory to indicate how these perceptual problems could be solved.

One of the modern experiments that has examined how a partial can be heard out from a complex tone was carried out by Leo van Noorden in the early 1970s.[291] He showed that if we listen to a rapid alternation of a pure tone A and a complex tone C that contains A as one of its components, and if A is presented at the right intensity on every cycle of the AC pattern, we will hear A occurring twice on each cycle (we will discuss the intensity issue later). That is, A will appear to repeat with twice the tempo of C. This experiment demonstrates only one-half of the old-plus-new heuristic. The component of C that matches the frequency of A is indeed being extracted from C as a separate sound; we know this because we can hear it. But because this component is such a small part of the harmonic structure of C (only 1 partial out of 10) we are not able to hear whether the remainder has a different quality after the extraction of this component.

Another fact about the old-plus-new heuristic is also visible in van Noorden's results. The capturing of the embedded harmonic improves with a shorter interval between the pure tone and the complex tone. It became stronger as the rate of alternation was sped up from 91 msec to 63 msec onset-to-onset time (for tones of a fixed duration of 40 msec). This dependence on temporal proximity operates in the same way here as in the streaming illusion that occurs with rapid alternations of high and low tones where the grouping of tones in the same frequency region gets stronger as the time delay between them is reduced. This suggests that the sequential integration is carried out by the same process in both cases.

The old-plus-new heuristic might be restated as follows: "Look for a continuation of what went before and then pay attention to what has been added to it." We must remember that the new thing that gets added might conceivably contain components at some of the same frequencies as the earlier sound. If the auditory system took away all the frequencies that matched those it had heard earlier it might take away too much. The later arriving tone might be left with no energy at the shared frequencies. Amazingly, the partitioning process seems to take away just enough and no more.

This was demonstrated in an experiment by Richard Warren and his colleagues.[292] The stimuli were noise bursts that were filtered to be one octave wide. Two bursts of noise of the same spectral content were alternated. The only difference between them was that one was more intense than the other. Under these conditions, the listener heard the less intense one as continuing through the more intense one. That is, the less intense one appeared to be present all the time, with the more intense one joining it at regular intervals. One strange effect occurred when the intensity difference between the two alternating sounds was less than 3 dB, for example, when an 80-dB sound alternated with one at 82 dB. The "added" sound (the one heard when the 82-dB burst came on) seemed less loud than the "constant" one (at 80 dB). This can be explained by remembering that because of the nature of the decibel scale, the decibel values of two sounds cannot simply be added together when the sounds are added. In the case of noise bursts, when two bursts of equal intensity are added together they make a sound that is just 3 dB louder than each of the originals. The auditory system was apparently mindful of this fact. Because of the similarity of spectral content, it thought that the more intense noise was a continuation of the less intense noise with an additional noise added to it. Then it had to calculate how intense the "added" noise must have been to augment the total intensity by 2 dB. It concluded that the added noise could not have been as intense as the constant noise since the increase was less than 3 dB. This explanation is supported by the results from another condition in which the stronger of the two alternating noises was exactly 3 dB more intense than the weaker one. In this case, the auditory system heard a constant noise joined periodically by another that was equal to it in intensity. The arithmetic of the auditory system seemed to be fairly precise.

I have used words such as "thinking" and "concluding" metaphorically to describe what the auditory system does under these circumstances. It is a metaphor because the process being carried out by the auditory system differs from ordinary thought in that we are not conscious of any inference process going on. Helmholtz called it unconscious inference and we can do so too, as long as we realize that the metaphor that we are using is based only on the results of the process and not on its detailed inner nature, which may or may not work like a thinking process. It is best, perhaps, to think of it as a process that has evolved in our nervous systems to deal with problems in the accurate use of sensory information. To do so, its information-processing rules must respect certain physical laws that

govern how our senses receive the information. In audition, this means respecting the laws of mixture of sound.

To respect a law does not necessarily mean having that law explicitly encoded in the neural circuitry that processes the information. It simply means that the process should deal effectively (by whatever means) with a world that is structured by these physical laws. To do so may sometimes requires the neural process to have distinct parts that mirror distinct physical laws or the relations imposed by those laws on the information that reaches our senses. This would allow the neural system to somehow model the interaction of a number of physical laws as they shape the evidence that strikes our sense organs. I have dealt with the issue of the composition of physical effects on sense data in my earlier writing.[293] The issue of the complementarity of perceptual processes to physical laws has also been dealt with by Roger Shepard.[294]

Before drawing conclusions about what the experiment on alternating bursts of noise means, we have to remember that the system, although not really faced with a mixture of sounds, was acting as if it were. The spectral similarity of the two bursts of noise pointed to the conclusion that the first burst continued through the second. After this conclusion was reached, the auditory system showed that is was able to deal with the physics of mixtures by extracting from a mixture just the amount of energy that might have been contributed by a continuing sound and by using the residual energy to build a description of an added sound.

This experiment on alternating noise bursts seems to indicate that the auditory system makes just the right computation about the components of mixtures. However, a second experiment points to a limitation of the ability of the auditory system to do so. Van Noorden found this limitation when he alternated a pure tone, A, with a complex tone, B, to get A to capture one of B's harmonics into a sequential stream. In his experiment, the best capturing did not occur when the intensity of A was the same as that of the harmonic it was capturing. The best intensity for A depended on the degree to which the target harmonic was masked by the other harmonics of B.[295]

The factor of masking showed itself in the following ways: The higher the harmonic he was trying to capture, the lower he had to set the intensity of A. We know that the higher components of a mixture are masked by the lower ones more than the lower ones are by the higher. It seems, then, that the intensity of A was required to match the "sensory strength" of the target as it existed after masking, not its physical intensity. This conclusion is supported by the finding, in the same experiment, that if van Noorden removed some of the other

harmonics from the mixture, reducing the masking of the target harmonic, he had to make the capturing tone louder.

How does this finding relate to the observation that the auditory system knows what happens to sounds in a mixture? It seems that it does not know enough about masking to match the original physical levels of A and the target harmonic. If A had really continued into the mixture, we would expect it to remain at the same intensity, and therefore the auditory system should find the best match between a target and a captor of the same physical intensity. However, it did not. I have tried out a number of explanations for the masking findings that employ the assumption that the nervous system knows about masking and tries to cancel its effects, but they all fail to explain the results of this experiment.

These results, then, show less ability on the part of the auditory system to arrive at the true facts about a mixture than did the earlier experiment on the alternation of louder and softer noise bursts. But there is a difference in what was being asked in these two experiments. In the case of the alternation of noise bursts, the question was not whether A (the the less intense burst) would be factored out of burst B; it was how much of B would be left behind. That was not the question being asked in van Noorden's experiment. In it, there was a competition between factors that favored taking the harmonic out (sequential similarity) and those that favored leaving it in (harmonicity and simultaneous onset). The experiment tried to find the best way to capture the component. So the two experiments are not exactly comparable. Perhaps the apparent knowledge that the auditory system has of the laws of physical mixture and of masking depends on the task and on the stimuli. This may be because the design of the auditory system has been optimized for several factors, not only the ability to deal with mixtures.

Spectral Relations

So far we have seen one way in which the auditory system deals with the problem of grouping the simultaneously received sensory components into sets that represent different sound sources. The old-plus-new heuristic looks at the sequence of events to find continuing ones that are hidden in a mixture. Other methods for deciding on the grouping of components do so without looking at earlier events. They group simultaneous components at different frequency regions by a consideration of certain relations between them. The next family of methods to be discussed is concerned with spectral properties that relate the different simultaneous components.

Properties of the Harmonics

All other things being equal, it appears likely that the further apart in frequency two simultaneous components are, the less likely they are to fuse. I recall once, when listening to an air conditioner, hearing two distinct noisy sounds, a low one and a high one. This prompted me to wonder whether I was hearing two sounds because there was a concentration of energy in two parts of the spectrum with a gap in between. It seemed to me that the auditory system might find some utility in segregating disconnected regions of the spectrum if it were true in some probabilistic way that the spectra that the human cares about tend to be smoothly continuous rather than bunched into isolated spectral bands. Since then I have collected a number of experimental observations, demonstrations, and theoretical speculations that bear on this issue.

The first has to do with the perceptual isolation of harmonics in a complex tone. When we listen to a complex tone in a nonanalytic way it seems that the first two partials are well integrated perceptually with one another and with all the rest of the harmonics. But are they really? We know that the harmonic series hangs together to generate the percept of a single tone, but are all the harmonics equally strongly bound into the fused mass? It has often been reported that it is easier for a listener to hear out the fundamental and the lowest partials of a complex tone.[296] A possible explanation for this finding depends on one of the properties of the peripheral auditory system.[297] Harmonics in sounds are linearly spaced in frequency whereas the mapping of frequency onto place on the basilar membrane is logarithmic. In log-frequency units the lower harmonics are spaced further apart from one another than the higher harmonics are. The way to tell whether this spacing is important is through experiments in which the harmonic in question is captured by a preceding pure tone of the same frequency into a sequential stream. Van Noorden used a stimulus in which a rich tone with 10 harmonics was alternated with a pure tone and supplied three pieces of evidence on this question. The first is that the lower harmonics were easier to capture out of the complex than the higher ones were. This was explained by saying that their representations on the basilar membrane are further apart. Second, it was easier to capture the components from a spectrum that contained only the odd harmonics than from a spectrum that contained consecutive ones. Again the explanation is that in the odd-harmonic spectrum the harmonics are twice as far apart as in the consecutive-harmonic spectrum. Third, a harmonic was easier to capture out of the complex tone when neighboring harmonics were removed. We can conclude that the greater the frequency separation between a harmonic and its

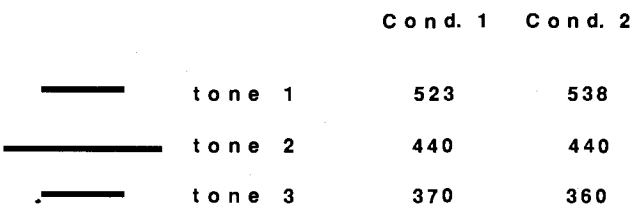


Figure 3.5
A tone passing through a "field" of other partials. Left: diagram of the stimulus pattern. Right: the frequency for each of them in two conditions. (After Vicario 1982.)

nearest frequency neighbors, the easier it was to capture it out of the complex tone.[298]

There is another effect, reported by Giovanni Vicario, that is consistent with van Noorden's findings.[299] It was obtained using the pattern of tones shown at the left in figure 3.5. A pure tone (tone 2) was sounded alone briefly and was then joined by another set of partials (tones 1 and 3); the accompanying partials were then shut off and the tone continued briefly by itself. In some cases, tone 2 could be heard as passing right through the rich tone generated by the field of partials and in others it did not pass through but seemed to be replaced by the rich tone and then to reappear afterward. Vicario found that the perceptibility of tone 2 was not reduced by increasing the number of tones that it had to pass through. He added four more, two higher than tone 1 (622 and 740 Hz) and two lower than tone 3 (311 and 262 Hz). The lack of any effect suggested that it was only the local proximity of tones 1 and 3 to tone 2 that made tone 2 hard to hear. The result is suggestive rather than conclusive. It may be that the immediate neighbors of tone 2 had some special harmonic relations to it as well as being closer to it in frequency. However, the suggestion that nearest neighbors exert the strongest effects in absorbing a tone into a complex is consistent with van Noorden's observations.

The basilar-membrane explanation for these frequency proximity effects has nothing to do with the utility of these effects for scene analysis. If the nervous system cannot resolve simultaneous tones that are close in frequency it will not be able to put them into separate streams no matter how desirable this might be. Yet I do not believe that an inability to resolve the separate harmonics can be the entire explanation for the greater tendency of the auditory system to fuse partials that are nearer in frequency. I grant that it could apply to tones that were very close in frequency, let us say within a critical band of each other (on the order of three or four semitones). How-

ever, it cannot apply to a tendency observed in some other experiments. These experiments used patterns like the one used by Bregman and Pinker and shown in figure 1.16 of chapter 1. The only difference was that each of the two simultaneous tones, B and C, was an amplitude modulated tone. The experiments on the perceptual fusion of such sounds will be discussed in more detail later. For the moment, I want to focus on one finding. Even when the partials are as much as 14 semitones apart, they tend to fuse more strongly than partials that are separated by 16 semitones.[300] With such large frequency separations between pairs, there is no question of whether the nervous system is capable of resolving them as separate sounds. The preference for fusing sounds that are nearer in frequency must have some other basis and may well be "designed in" for scene-analysis purposes rather than being a by-product of a failure to resolve parts of the spectrum.

When Pinker and I found that a partial could be captured out of a mixture we were using a mixture of two partials separated by about an octave (see figure 1.16).[301] Although an octave may seem large, these partials were no more separated than the first two harmonics of a harmonic series, f and 2f, are. Yet, in informal experiments, I have found that the second harmonic of a complex tone that contains many harmonics is harder to capture than the higher partial in the two-partial complex. Plomp has also observed that the ability of listeners to identify a partial in a two-partial complex tone was very much better than in a tone containing more partials. The frequency separation between adjacent partials that permitted perceptual isolation of the partial was as much as three times smaller in the case of the two-partial tone.[302]

What of the observation that higher harmonics are harder to capture out of a complex tone than the lower ones? Can this effect be explained by the greater crowding of the representations of higher partials on the basilar membrane? An experiment done by Gary Dannenbring and myself bears on this question.[303] It employed a complex tone whose three partials were logarithmically spaced in frequency and hence equally spaced on the basilar membrane. The stimulus pattern consisted of the rapid alternation of a pure tone with this complex tone. In different conditions, the pure tone was set to the frequency of one of the harmonics of the complex tone and therefore acted as a captor that tended to capture its target harmonic out of the stream and into a sequential stream that included both the captor and the target. The three components of the complex tone were spaced by octaves, at f, 2f, and 4f. Even with this logarithmic spacing, the higher the harmonic, the harder it was to capture.

Therefore an explanation in terms of basilar-membrane spacing was not adequate.

However, there is another high-low asymmetry in the auditory system that may be used to explain this result. This is the well known "upward spread of masking." It is found, using a variety of methods, that lower tones mask higher tones better than higher ones mask lower.[304] This could explain why the higher partials are hard to extract from the complex tone. Remember that the ease of extracting them is judged either by just listening for them or else listening for them in the sequential stream into which they are pulled out by a captor tone. In both cases, since they are less audible, they may appear harder to capture. Two experimental results seem to support this explanation. The first is the observation that if you want to capture a harmonic out of a complex by alternating the complex with a captor tone, you get the best results when you make the captor softer when trying to capture a higher harmonic.[305] This seems counterintuitive; you would think that a louder captor would be better. However, we must remember the manner in which the experiment is usually carried out (for example, by van Noorden). The listener hears the rapid alternation of a pure tone with a complex tone and listens for a double beating of the pure tone, that is, for the pure tone appearing twice as often as the complex tone. This double beating means that the target harmonic is being heard as a sound separate from the complex tone, so that there are two pure tones (captor and target harmonic) for every repetition of the complex tone. Now suppose that due to upward spread of masking the target partial is registered very weakly by the auditory system. Then even if it were pulled out of the complex by the captor the pure-tone stream would contain a strong tone (the captor) alternating with a weak tone (the target) and the listener might be affected by the contrast between a strong and a weak tone and be led to believe that the weak tone was just not there. On the other hand if the physical intensity of the captor were lowered so as to achieve an equality of perceived intensity between the captor and the target the listener would judge that both were present in the pure tone stream. This masking-based explanation of the difficulty of hearing out the higher harmonics is also supported by the following finding: If you raise the intensity of the target harmonic it is easier to pull out of the complex, and this effect is more pronounced for the higher partials.[306] This would be expected if the perception of the higher partials were especially hurt by masking and if raising the intensity helped to overcome the masking.

Dannenbring and I found that the intensity pattern in which higher harmonics were hardest to capture was one in which a higher harmo-

nic was weaker than a lower one. Despite the ready physiological explanations that come to mind to explain this fact, we should not neglect its implications for scene analysis. Most natural pitch-like sounds, such as the vowels of the human voice or most instrumental tones, have a long-term spectrum in which the higher partials are weaker than the lower ones. Sometimes this pattern is temporarily violated when strong resonances are imposed on the spectrum, as happens when the human speech apparatus forms the resonances that define one particular vowel, but if we average over many specific examples, the spectrum does have the property of a falling intensity with a rising frequency (at least in the range above 300 Hz).[307] If the auditory system prefers to accept as parts of the same sound, partials that fall off in intensity with increasing frequency, this will give a statistical boost to the grouping process and increase the likelihood of fusing partials that have arisen from the same acoustic event. The fact that this may be the result of a capacity limitation at the physiological level may be another example of the point that I made earlier—a functional accomplishment may be subserved by a physiological breakdown. Indeed, this point encourages us to ask the question "Why is the auditory system arranged so that lower partials mask higher ones? Does this arrangement serve some purpose?"

It also appears that the smoothness of the spectrum is important and that harmonics that are raised in intensity will segregate more readily from the others.[308] This can be understood in terms of the masking of one tone by another. A louder component is less likely to be masked and more likely to mask the others that accompany it. It is hard to find a direct explanation, in scene-analysis terms, for the segregation of more intense components from weaker ones that accompany them. Are they less likely to be parts of the same sound? Perhaps this basis for segregation is a by-product of the evolved design of the system and not specifically evolved to serve a function of its own.

Harmonic Relations (Harmonicity) in Complex Tones

There is a particular relation that can hold among a simultaneously present set of partials. They can all be harmonics of the same fundamental. There is a good deal of evidence to suggest that if they are, they will tend to be assigned to the same stream; that is, they will be fused and heard as a single sound. Let us call this the "harmonicity" principle.

There are a number of facts that point in this direction. One is the simple observation that a set of harmonics arising from a single fun-

damental is not heard as a large number of individual partials but as a single sound. This has been noticed over and over again.[309] While this placing of related partials into a single package serves to group components that have come from the same source there is a more definite purpose as well—to calculate the pitch of the fundamental. Most objects, when they are set into vibration, move not just with a single vibration frequency but with a large number of frequencies. In many cases, including the human voice and all instruments with clear pitches, the frequencies of vibration are all multiples of the lowest frequency (the fundamental). In such cases, the different frequencies are called harmonics. If the lowest frequency is called f, the harmonics have the frequencies f, 2f, 3f, 4f, ... Our auditory system seems to have evolved a method of finding out what the fundamental is, even if only some of the higher harmonics are present, with the result that the pitch we hear is the pitch of that fundamental. Even if we filter out the fundamental frequencies of the successive notes in a melody, as for example when we play it over the telephone or any equivalent low-fidelity reproduction equipment, we still hear the melody with its pitches unchanged despite a change in the fullness of the sound. How can we be hearing the pitch of the fundamental even though the fundamental has been filtered out? This question has been referred to as the mystery of the missing fundamental. How can the auditory system hear a sound that is not there? Apparently it is capable of making use of the regular spacing between the existing harmonics to determine the pitch of the missing fundamental.

It is important to the auditory system to group the harmonics together into the same analysis. This will seem to be a strange statement to those hearing scientists who are used to thinking of the typical laboratory setup in which only a single set of harmonics is played to the listener at any one time. In this case the issue of grouping seems meaningless and it seems sufficient to say that the listener just finds a fundamental that accounts for all the components that are heard. But the situation changes as soon as two complex tones are played at the same time. If the two fundamental frequencies are unrelated, an analysis that tries to find a single fundamental for all the partials that are present will fail. Yet we know through listening to music that we can hear two or more pitches at the same time. There has to be a way to base the computation of each component pitch on only a subset of the partials. Assuming that this is somehow done, what we hear in the presence of a mixture of two harmonic series (as when two musical tones are played at the same time) is not a large set of partials but two unitary sounds with distinct pitches.

Models of the Pitch-Analysis Process

Because of the intimate connection that probably exists between the computation of pitch in the auditory system and the perceptual grouping (fusion) of simultaneous partials, it is necessary to discuss theories of pitch perception. I cannot describe all the theories that have been proposed to describe how the auditory system calculates the fundamental from a set of harmonics. They are discussed in many textbooks.[310] I want, however, to discuss some of the issues in pitch perception.

There are two phenomena that current theories of pitch have tried to explain. The first is how we can hear the pitch of the fundamental even when it is filtered out of the spectrum. The second is how we derive a pitch for a spectrum whose components are not exactly harmonically related. I have already described how one can physically create a spectrum in which the fundamental is missing. Let me now describe how an inharmonic spectrum gives rise nonetheless to a pitch.

Due to the genius of modern sound-producing equipment, it is possible to present a listener with a tone formed of partials that are inharmonic. That is, they do not fall into a neat harmonic series. Nonetheless, they often give rise to a pitch that is not simply the pitch of one of the partials that is present. Apparently this happens because our mechanism for deriving a pitch from a set of partials goes to work on this inharmonic array of partials and comes up with an answer which is simply the best it can do with this weird input. It should be noted, however, that the pitch sensation derived from an inharmonic set of partials is never as strong and clear as the one derived from an equal number of harmonic partials.[311] Furthermore, the sound that one of these sets gives rise to is not as perceptually unified as the one that a harmonic series provides. Much research has been carried out on the perception of the pitch of inharmonic partials.

One way to derive inharmonic partials is to start with a set of harmonic ones and then shift their frequencies up or down by adding or subtracting the same constant to each. For instance, let us take some of the harmonics related to a fundamental of 100 Hz, for example 300, 400, and 500 Hz. These are all multiples of 100 Hz and they will yield a pitch related to 100 Hz. Now let us add 7 Hz to each frequency so that they are 307, 407, and 507 Hz. Notice that they are no longer multiples of any fundamental frequency (except 1 Hz, which is too low to be heard as a pitch). Despite the inharmonicity of the shifted group of partials we will still hear a pitch. The pitch will be just a bit higher than the 100-Hz pitch that we heard before we added the 7 Hz.[312] At the same time we may hear the pitches of the separate par-

tials. Apparently in this case inharmonic tones are accepted into a pitch analysis, at least partly. While they may not lose their individual identities entirely, they contribute to a global estimate of pitch. This pitch experience is, however, weaker than the one that occurs with a harmonic series. It appears, therefore, that the admitting of partials into a pitch computation is not all or none. They can be weakly included as in the present example.

Theories of Pitch Analysis Different theories of auditory pitch analysis have addressed two problems: How we can hear the pitch of the missing fundamental in a harmonic series, and how we can hear a pitch in shifted sets of partials. The theories tend to differ on two dimensions: (1) whether they see pitch analysis as based primarily on "place" information or on "periodicity" information, and (2) what method is used to derive the pitch from the type of information that is used.

Some theories use the fact that different places on the basilar membrane of the inner ear respond maximally to different pitches. Therefore, if the auditory system knows what places on the basilar membrane are responding, and with what amplitude, it can infer a lot about the spectrum of the stimulating sound. This is the "place" information. Other theories use the fact that the part of the basilar membrane that responds best to a given frequency component also tends to vibrate at the frequency of that component. Therefore there is information in the vibration rates (or periodicity) at each position along the basilar membrane, and the auditory system could use it to infer the spectrum of the stimulating sound. This is the "periodicity" information. A neural recording of this periodicity is provided by the fact that each small region of the basilar membrane has its own nerve supply that primarily registers the frequency to which that portion of the basilar membrane is tuned. If the local region of the basilar membrane with its associated nerve supply is thought of as a channel, we can see that the output of this channel behaves like a band-pass filter. The filter, however, does not completely exclude nearby frequencies, especially those that are lower than the tuned frequency. Therefore the neural record of each frequency in the incoming signal is perturbed by the existence of other frequencies.

So far we have described how the theories differ in their description of how the spectrum of the incoming signal is estimated by the auditory system. They also differ in how they think that the estimates are used to derive a pitch. The methods appearing in different theories all try to derive the fundamental frequency of the incoming spectrum, but they go about it in different ways and therefore are fooled in

different ways when an inharmonic spectrum is encountered. The most direct way of using the spectrum is by a calculation of the fundamental to which the received partials are all related harmonically. One of the theories assumes that there can be some error in the original estimation of the frequencies of the partials and that the system overcomes this problem by doing a "best fit" calculation on the set of estimated frequencies.[313] Models of this type have been referred to as pattern-recognition models because they require the auditory system to recognize the incoming pattern of partials as a set of harmonics that is appropriate for some particular fundamental.[314] That is how they solve the problem of the missing fundamental, the first phenomenon that was a challenge for modern-day theories of pitch. They explain the second phenomenon, the pitch of an inharmonic spectrum, by assuming that the auditory system still treats the partials as if they had arisen from a common source and tries to find the fundamental of which they are harmonics. Because the partials are really not harmonically related, the fundamental that is found will not fit very well, but the system will put out the pitch of this tone as its best guess.

The second method for deriving the fundamental does not directly use the information about the frequencies of the detected partials, but uses information about the beats between them. To understand this theory it is important to remember that equal steps along the basilar membrane do not represent equal steps in frequency but equal steps in log frequency. Octaves (doublings in frequency), for example, are equal steps apart. Therefore the harmonics, because they are equally spaced in frequency rather than log frequency, will stimulate a set of tuned basilar-membrane locations that are crowded closer and closer together as the harmonic number goes up.

The neural outputs of an individual basilar membrane location is not entirely driven by the single frequency to which that location is tuned, but by adjacent frequencies as well. For a harmonic series, this would happen more for higher harmonics because they are more crowded on the basilar membrane. Therefore, if you could look at the neural output of the basilar membrane, in its lower channels you could see firing rates corresponding to individual harmonics; in higher channels you would mainly see firing rates caused by the interaction of two or more harmonics. Any physical system (such as a band-pass filter) faced with unresolvable combinations of frequencies will register beats (periodic fluctuations in the amplitude envelope) whose rate will correspond to the differences in frequencies between pairs of harmonics. Recall that in a harmonic series the harmonics of a fundamental frequency, f, are f, 2f, 3f, 4f, ... Notice that their frequencies go up in increments of f. That is, they are spaced from their

nearest neighbors by f, the frequency of the fundamental, and from their nonnearest neighbors by multiples of f. When a set of harmonics is not resolvable, the beat rate will be dominated by f and have multiples of f as well. This means that the interference pattern between adjacent harmonics will be repeated at a regular period. Conveniently, this period is the period of the fundamental. This will hold independently in every spectral band in which there are unresolvable harmonics. In these frequency regions, in the higher parts of the spectrum, while the neural mechanism is not fast enough to follow the repetitions of the waveforms of the individual harmonics it is fast enough to follow these beats. Therefore, there can be an independent estimate of f in a number of spectral bands in the higher parts of the spectrum. In the lower parts, where the auditory critical band is narrower and the harmonics are fully resolvable, the harmonics themselves are of a low enough frequency that their individual periods can be registered by a neural mechanism. It is, in principle, possible for a neural mechanism to use both the beat rate in the upper part of the spectrum and the rate of the separable harmonics in the lower part.

According to the proponents of "temporal" theories, it is totally unnecessary to compute the fundamental from the pattern of frequencies of the incoming fundamentals. The interference pattern in each basilar membrane channel does it for us. That is the explanation of how we hear the missing fundamental. The explanation of the pitch of inharmonic complexes is derived in the same way. Let us suppose that we have taken a group of three harmonics, 500, 600, and 700 Hz, that would normally be sufficient to specify a missing fundamental of 100 Hz and shift them all up by 15 Hz so that the frequencies are now 515, 615, and 715 Hz. In contrast to the case of harmonically related partials, the interference patterns for these shifted partials will not be exactly periodic and therefore the process whose job it is to detect the periodicity in the interaction of harmonics will find a pitch that is partly related to the 100-Hz separation between them, but which is a bit higher due to the nonexact repetition of the interference pattern.

Listening to Inharmonic Partials

As I have already pointed out, sounds that are composed of inharmonic partials give a weak sense of pitch and do not fuse very well.

For example, an inharmonic set of partials can be derived by simply mathematically stretching the separations between the partials of a harmonic series. Consider the ratio of 2. In a harmonic series, this is the separation between the frequencies of the first and second harmonics, or the second and fourth, or the third and the sixth, and so on.

By stretching all the frequency separations between the partials (on a log scale) so that the frequency ratio that relates these particular pairs of harmonics is now 2.2 instead of 2, a new set of frequencies can be derived for all the partials so that they are no longer harmonically related. The partials are now only 10 percent further apart than they were before (for octave relations) but they now no longer are heard as a single coherent sound. Instead, the lower partials are often heard as individual tones and the higher ones remain undistinguished from one another as a buzzing tonal mass. The sense of pitch is very indistinct. The judged fusion seems to continuously get worse when we either gradually stretch the spacing of the partials from what it is in a harmonic series to 7 percent further apart or shrink their spacing gradually to 7 percent closer together (for octave relations).[315] John Pierce, who has a lot of experience in listening to such sounds, has also noticed that when two stretched series of partials are sounded at the same time, you do not hear only two distinct sounds as you do when listening to two harmonic sounds.[316]

Clearly two things happen with the harmonic partials that do not with the inharmonic ones: the set of partials is heard as a single coherent tone, and that tone has a clear pitch. Therefore we see differences both in perceptual unification and in pitch extraction. A natural question comes to mind: is it the same process that extracts pitch and that unifies the partials into a single sound? Although we do not know the answer, we can remark that both accomplishments require the auditory system to group acoustic components. There have been recent attempts to extend the theory of how the human auditory system does pitch extraction so that it can explain how it is possible to hear more than one pitch at a time. Briefly, the older single-pitch theory supposed that the auditory system had some way of examining the neural impulses coming out of the inner ear and calculating the fundamental frequency that would best explain this neural output. The newer approach extends the method so that it finds the two best fitting fundamentals.[317] Undoubtedly, the approach will be further generalized to find the number of pitches that the human ear can hear, a number greater than two in many listening circumstances. Any theory of this type would group the separate sets of harmonics in the course of finding the separate fundamentals.

We have looked at the case in which all the partials are inharmonically related to one another. What happens if we start with a tone that is built from a harmonic series and then mistune just one partial from the rest of the partials? It has been reported that this partial becomes easier to hear out from the complex tone.[318]

Here is a description of what it sounds like when a harmonic partial is gradually mistuned:

> When a low harmonic component is mistuned by a small amount, less than 1%, the pitch of the complex as a whole shifts a little, but the tone still sounds like a harmonic complex, and the mistuned partial is not heard as a separate tone. When the mistuning is increased to 1%–1.5%, the pitch of the complex shifts further, and the complex can now just be distinguished from a truly harmonic complex. However, the mistuned partial is usually not heard as a separate tone. If the mistuning is increased to 1.5%–3%, the pitch of the complex shifts still further, but the mistuned partial is now just heard as standing out from the complex. If the mistuning is increased to 6%, the pitch shift of the complex generally declines, and the mistuned partial is heard even more clearly as standing out from the complex.[319]

Basically what happens is that at first the mistuning of the tone changes the pitch of the complex tone. This is consistent with the idea that some pitch-determining mechanism is trying to estimate the fundamental of the set of partials and the mistuning of one of them throws its calculations off. With greater mistuning the partial is somehow excluded from the pitch computation and the pitch returns to what it would have been if the partial had not been present. This intimate connection between the calculation of pitch and the decision about which tones go together suggests the existence of a single system whose job it is to carry out both tasks and that uses the harmonicity of the set of partials to make its decisions.

Logically, either group of theories, those that depend on recognizing a pattern of partials indicative of a certain fundamental or those that depend on the interference pattern within frequency-specific neural channels, could explain how the pitch system could exclude some partials from a global pitch analysis. A system using the pattern-recognition approach might be able to look for the set of harmonics related to every possible fundamental (within a certain range of possibilities) at the same time. For convenience we can imagine the system as possessing a set of harmonic-series templates, each one specific to a particular fundamental, which it applies to the pattern of active frequency channels. For example, the template for the 100-Hz fundamental would look for activity at 100, 200, 300, ... If a sufficient number of these were showing activity, it would register a global pitch at 100 Hz. Providing it could tell which partials had contributed to the decision, it might remove the partial pitches of these components from perception, leaving only the partial pitches of the

badly fitting partials that had not activated the template. The method can be thought of as a "harmonic sieve," which allows through its holes only those tones related to the template and stops ones that are too far away from the frequencies of the acceptable harmonics.[320]

If, on the other hand, our auditory system estimates the fundamental from the repetition rate of within-channel interference patterns, it would have to "sift out" nonharmonics in a different way. It could look for neural channels whose periodicities were related (or almost so) and accept only those similar channels into the same pitch analysis. The remainder would be excluded.[321] There is some reason to believe that it would have to do this even if it were not interested in partitioning the spectrum to assist later pattern recognition. If it allowed badly fitting channels into the pitch estimation process, then, when it was estimating the pitch of a human voice, for example, it could come up with an estimate that was really a mixture of the effects of two different voices and hence not the pitch of anything at all.

We see then that any theory of pitch computation that can compute the pitch of a sound and not get confused by other co-occurring sounds would be, in effect, doing a scene analysis based on the harmonicity principle. That is why I said earlier that pitch estimation and sound separation were so intimately linked.

We should recall again at this point what the examples and theories mean for scene analysis. When a mixture of sounds is encountered, it may be useful to analyze the mixture for the existence of one or more good harmonic series. This strategy is useful since in the most important sound to us, the human voice, the partials come very close to forming a harmonic series. If we are able to detect a harmonic series, we can determine its fundamental frequency, assign a pitch to it, and then remove all the harmonics related to this pitch from the set of partials still being considered. Then the remainder can be taken alone and subjected to further analysis to discover other groups of partials that go together.

One consequence of the exclusion of a partial from a fusion group is that it should become freer to enter into sequential streams with earlier sounds that resemble it. We saw an example of this in the experiment of Bregman and Pinker where the upper tone, B, of a two-tone complex, BC, was freer to group with a preceding pure tone when the conditions were less favorable to the fusion of B and C (see figure 1.16). We should find, then, that inharmonic partials will more readily be captured away from their simultaneous companions and into sequential streams. This property would make it more likely that a sound that was suddenly joined by another that was unrelated

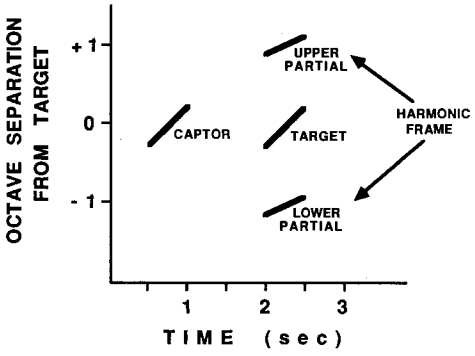


Figure 3.6
Capturing the central glide in a three-component glide. (After Bregman and Doehring 1984.)

to it would be more easily grouped with its own earlier part than with its accidental companion.

There is evidence both for and against this prediction. In one of Bregman and Pinker's experiments, they varied the relation between B and C so that they either formed ratios of 3:2, 2:1, or 3:1 or else were mistuned by a semitone from each of these simple ratios. They found no evidence that the harmonicity played a role in the ability of A to capture B into a sequential organization. Similarly, Plomp found it possible to "hear out" unrelated tones as easily as harmonic tones from a tonal complex.[322]

On the other hand, harmonicity did play a role in another experiment that was very much like the preceding one except that the component tones were all sinusoids gliding in a straight line on log-frequency-by-time coordinates as in figure 3.6.[323] A pure tone glide alternated repeatedly with a complex of three simultaneous glides and had the same frequency and slope as the middle of the three simultaneous glides so that in some conditions it captured the middle one into a sequential stream. Now in one set of conditions, the three simultaneous glides were all parallel on the log scale (unlike the case shown in figure 3.6 where the "target" glide does not move in parallel with the glides of the "harmonic frame"). In the parallel conditions, the glides remained in the same frequency ratio to one another throughout their gliding paths. In some of these conditions, the three glides were spaced by octaves, forming the ratios 4:2:1 throughout (the octave relation is a very powerful one for fusing tones). In other conditions, the two outer glides, the highest and the lowest, remained in the two-octave (4:1) relation as before, but the middle glide was

mistuned from the octave relation by a quarter octave so that now the ratios between the glides were 4:1.68:1. Under these conditions, where the middle glide was mistuned from the other two, it was much easier to capture it into a sequential stream.

The discrepancy between the results of the two experiments cannot be attributed to the use of glides in the second because in that experiment there was a condition in which the slopes of the glides was 0 octaves per second (in other words, they were constant-frequency tones). This condition showed the same effects of harmonicity as the other ones. The only important difference was that the glide experiment used three simultaneous components instead of two. I think this was the crucial difference. In the inharmonic condition the outer two glides defined a harmonic frame into which the middle tone did not fit. Apparently it is not sufficient to have only a single tone define a potential harmonic series into which the other does, or does not, fit. I suspect that in the Bregman-Pinker experiment there was no real manipulation of harmonicity because there were too few partials to define it. If I am right, we should find that increasing the number of partials that maintain a harmonic relation to one another should increase the isolation of a partial that is mistuned from this group. To my knowledge, there has been no definitive study of this sort.

One of the surprising things about inharmonic complexes is that they have pitches. Somehow whatever process normally computes the pitch of harmonic complexes is activated sufficiently to generate a pitch. However, when we listen to an inharmonic complex we are often most aware of the pitches of certain individual partials that are not fusing with the others. We need a vocabulary for talking about different kinds of pitches. I would propose that we assign the name "partial" pitch to the pitch of an individual partial and the name "global pitch" to the pitch that is computed on the basis of a set of partials. The first thing to notice when we listen to an inharmonic complex is that the global pitch is weak. Partial pitches are more strongly heard. In harmonic complexes, this is rarely the case. How can this be explained?

First of all, we know that the auditory system can compute more than one pitch at a time, otherwise we would not be able to hear the individual pitches in a musical chord. The fact that we hear both the pitches of partials as well as a global pitch in inharmonic complexes is simply another example of the fact that more than one pitch can be calculated at once. Furthermore, it seems plausible that the frequencies of the partials that we are hearing as partial pitches are contributing to the computation of the global pitch as well. Why, then, are the partial pitches heard with such difficulty in complexes that

contain only harmonic partials? The harmonics can be heard out by most people only if a pure tone, at the same frequency as the harmonic that they want to hear, precedes the complex tone.

Two explanations come to mind. The first is that we can hear out only the nonintegrated partials, those partials that are not being integrated to yield a global pitch. This would explain why we do not hear the partial pitches in harmonic tones: none of the partials is being rejected from the pitch analysis. This makes sense of the findings that I discussed earlier about the experience of a listener when a single partial in a harmonic series is gradually mistuned. Before it is rejected from the global analysis, the global pitch changes since the latter is now incorporating discrepant information. However, after the partial is rejected two things happen. The global pitch returns to what it was before and the partial pitch of the mistuned partial is heard. The implication of this explanation is that no matter how you remove a partial or a set of them from a global pitch analysis, whether by mistuning or by sequential capturing, the global pitch will be altered. As far as I know, this prediction has never been tested.

There is a second possible explanation for the absence of partial pitches in a harmonic complex. They may be actively inhibited by the nervous system. It would work this way: first, there would be a detection process for the existence of a harmonic series. This process would "know" the frequencies of the partials that were harmonically related to every fundamental. If a sufficient number of harmonically related partials were present at the same time, the detector would signal the existence of a global pitch appropriate to the inferred fundamental. At the same time, it would remove the individual harmonically related partials from the pitch analysis. Once a global pitch had been formed, the auditory system would try to inhibit the independent partial pitches of just those harmonics that were multiples of the perceived fundamental, in order to favor the perception of the global pitch.

The same theory would explain why we can easily hear both a weak global pitch and other partial pitches in an inharmonic complex. The existence of the global pitch would occur because some subset of the inharmonic partials managed to weakly satisfy the requirements of the detection process (which might not be precisely tuned so that it could tolerate some error). However, since even that subset of partials would not provide very strong evidence for a harmonic series, the global pitch would be registered only weakly. Furthermore we might expect the attempt by the global process to fail in its attempt to inhibit the partial pitches. This failure might occur because the pitch process knew only the set of *harmonic* partials associated with each

computed global pitch. Even when it was fooled into registering a global pitch by an inharmonic complex, it would try to inhibit the partial pitches for the set of *harmonics* that normally trigger it. Since the partials that actually stimulated it would not be the ones that it knew how to inhibit, its inhibition would miss them and their partial pitches would be heard.

This theory has a nonintuitive prediction to make: If we started with an inharmonic complex and mixed it with a pure tone whose frequency we could adjust by means of a knob, and began to adjust the tone up and down in frequency, we would be able to hear the partial pitch of this pure tone. (We might, of course, need some assistance in hearing the partial. This could be provided by alternating an isolated copy of the pure tone with the mixture.) However, certain frequencies of this pure tone would be predicted to be harder to hear as partial pitches in the mixture. These would be at the frequencies that were harmonically related to the weak global pitch that was being heard at that moment, and therefore being actively inhibited by the pitch detector. Again, as far as I know, there are no experimental data to support or refute this prediction.

However, we do have some information that pertains to one presupposition of the theory. We need to be able to assume that the auditory system has the ability to direct inhibition very precisely to just that set of partials that is related to a particular fundamental. This assumption is supported by an observation reported to me by Quentin Summerfield. It is based on the earlier observation that if a person is exposed to a harmonic spectrum that has a shape to it (that is, where certain frequency regions are enhanced) and then is switched to a flat spectrum (where all frequency regions are equally intense), the flat spectrum will be heard as having a shape that is the complement of the one that was listened to first. You can use this effect to make a person hear a flat spectrum as if it were a vowel. First you present an inducing spectrum that has the complement shape to that of the desired vowel, intense at frequencies where that vowel is weak and weak at frequencies where the vowel is intense. A subsequently presented flat spectrum will sound like the vowel.[324] This is interpreted as resulting from a briefly persisting inhibition of those frequencies at which the inducing spectrum was intense.

The relevance of this phenomenon to our previous question becomes visible when we look at a fact discovered by Summerfield and his co-workers in the course of studying this way of inducing a vowel percept.[325] In the previous case the fundamental of the spectrum of the inducing stimulus and the vowel were identical, but now they arranged it so that the fundamental frequencies of the inducing stimu-

lus spectrum and the subsequent flat spectrum were different (that is, the spectra had different frequency components as well as different spectral shapes). In this case the induction of the vowel percept did not occur. This shows that the inhibition process is accurate enough to target only the partials that were present in the first tone.

We are left with two possible explanations of why we can easily hear partial pitches in an inharmonic complex and not in a harmonic complex. Either the heard pitches are available because they were never integrated into the computation of the global pitch or because despite being integrated they have not been properly inhibited by the global process. Whatever the mechanism, we can see a good scene-analyzing reason for their being perceptually available when they are not harmonically related to the dominant global pitch. The harmonicity heuristic has decided that they are not part of the sound that generated the global pitch and therefore should be available to be heard in their own right or as part of some other sound or sequence of sounds.

Fusion and Segregation of Simultaneous Complexes

We have seen that if we hear a simultaneous set of partials in which some partials do not fit into a harmonic series, they are not fused with the ones that do. But what about more complicated cases in which we hear a group of instruments playing at the same time or a group of singing voices? Musicians believe that a simultaneous set of tones will blend better if they are in a harmonious or consonant relation to one another. This is different from the cases that were described previously. The earlier discussion concerned the rejection of components that were not part of a single harmonic series. In the present case, there are a number of tones present, each with its own harmonic series.

To simplify the discussion, let us consider only two notes of different pitches played together. In each note, the partials relate to a common fundamental, but this fundamental is different for the two notes. This being the case, our earlier discussion would lead us to think that the two sets of partials would be allocated to two pitch analyses. This is true; in such a mixture we will usually hear two pitches. However, our discussion led us to believe that as long as there were two different pitches, the tones would always be perceived as equally distinct from one another. This goes against the beliefs of musicians who generally hold the opinion that some pitches blend better than others.

It seems that two tones blend better when more of their harmonics coincide. We can start with the example of two tones, spaced by an octave, where the higher tone will (by definition) have a fundamental that is double the frequency of the lower one. Take a concrete case

in which the fundamentals are at 100 and 200 Hz. The harmonics of the 100 Hz tone will be at 100, 200, 300, 400, 500, 600, 700, 800, . . . , whereas the harmonics of the higher one will be at 200, 400, 800, ... With these fundamentals, every harmonic of the high tone coincides in frequency with a frequency of the low one, and half the frequencies of the lower one appear in the higher one. This is the highest overlap that the harmonics of two different pitches can have. The only way tones can have more than 50 percent of their components in common is if both are at the same pitch.[326] It is probably not a coincidence that the perceptual fusion of tones is highest if they both have the same pitch and next highest if they are an octave apart.[327]

The blending of tones to create an experience of a single richer tone is clearly exemplified in the design of the pipe organ. This use of harmonic relations will be described in chapter 5 in a discussion of the role of scene analysis in music.

Another example of the principle that tones with simpler harmonic relations blend better was provided by a demonstration created by Giovanni Vicario.[328] First a pattern was demonstrated in which a 440-Hz complex tone was accompanied by two other complex tones for part of the time that it was on. This was shown earlier in figure 3.5. Tone 2 was 2.5 seconds long and it was accompanied by the other tones for its middle 1.5 sec. In a first condition, the fundamentals of these tones were spaced by a minor third, which is approximately the ratio 6:5. Such a spacing made the harmonics of the adjacent pairs of tones fall into a moderately simple harmonic relation with one another but the extreme tones (tones 1 and 3) fall into a more irregular relation. In this case, tone 2 seemed to partially disappear at that point in time at which the other tones were switched on. In the second condition tones 1 and 3 fell into the ratio 3:2, which put their harmonics into a fairly simple relation with one another; however, the middle tone was mistuned to both of the others. In this case it sounded more independent of the others and did not tend to disappear when the other two came on. This effect can be thought of in the following way: Because the part of tone 2 that co-occurred with tones 1 and 3 was rejected from an otherwise good harmonic grouping, it was more easily grouped with its own parts that extended outside the three-tone complex and was therefore heard as a continuous long tone.

Perhaps the reason for the fusion of related tones can be most easily understood by thinking not in terms of fusion but its opposite, segregation. We then see that two individual pitches are more prominent when the two fundamentals have few frequencies at which their harmonics coincide. If the pitch calculation system tends to eliminate a

harmonic from further consideration after it has used it in the calculation of a global pitch, shared harmonics might not serve to define two different pitches very well. An absence of coincidence would allow the pitch calculation system a clearer, more independent look at the harmonics of each fundamental, and hence would allow it to clearly see that there were two.[329]

An indication that the auditory system prefers to have independent evidence that two pitches exist comes from an experiment in which listeners were required to separate the global pitches generated by two simultaneous sets of harmonics, each related to a different (missing) fundamental. The task involved recognizing the musical harmonic interval between pairs of tones.[330] Recognition was particularly hard when the harmonics were close enough to beat with one another; it was not so bad when they exactly coincided because at least the beating was eliminated. It was easiest when the harmonics occupied completely nonoverlapping regions of the frequency spectrum. Besides showing that the auditory system is least confused when the evidence is spectrally clear and separate, this finding suggests that the computations that calculate the pitch of a complex tone operate independently in local regions of the frequency spectrum.

The question of how the auditory system separates simultaneous speech sounds has attracted a lot of attention in recent years. This topic will be dealt with more thoroughly in chapter 6. For the moment we might simply point out that many theories of voice separation, both those that have been devised to explain human performance data and those that have been used to program computers to separate concurrent speech signals, start by trying to find the set of frequency components present in a mixture and then attempt to divide the total set into subsets that are harmonics of different fundamentals. There have been demonstrations with synthesized speech sounds that show that the auditory system tends to fuse two sets of partials that are in different portions of the spectrum, even sets presented to different ears, when each set is formed of consecutive harmonics and the two sets of harmonics are related to the same fundamental frequency. When the two sets are related to different fundamentals, they are heard as two separate sounds. There is some controversy about whether this segregation affects the listener's recognition of what is being said, but I believe that it has been clearly shown to do so in certain cases.

In the preceding section, I have presented evidence that the pitch estimation system acts to group harmonically related partials. We might conclude that this grouping is then used to derive other properties of the now segregated partials. This description implies a one-

way transaction, the pitch system influencing the grouping system and not vice versa. However, this appears not to be true. There is evidence that the pitch that is calculated can depend on cues other than harmonicity, cues that we think of as operating outside the pitch system. These will be discussed more fully later when we get to those particular cues. To give a brief preview, however, we will see that when coherent frequency modulation is applied to a set of partials, it increases the likelihood of the calculation of a global pitch based on that set, whether the set is harmonic or not. This suggests that other factors that can affect the grouping of simultaneous components of a sound can affect the computation of global qualities of the subsets of sounds that are derived from this grouping. In chapter 5 we will see that even such basic properties of sound as the *dissonance* of a group of musical tones, which was thought to be based on a raw acoustic phenomenon, can be strongly reduced if the tones that are acoustically dissonant with one another are caused to be perceptually segregated from one another. We will be able to add dissonance to the list of basic qualities that are affected by the segregation (or fusion) of simultaneous auditory components.

This section has reviewed the role of spectral relations in the decomposition of mixtures of sounds. Many observations have suggested that one of the heuristics for simultaneous grouping of spectral components is whether or not they fit into the same harmonic series. The auditory system can detect more than one harmonic series at a time, each related to a different fundamental, and construct a number of pitches. This pitch generation seems to go along with a merging of the identities of the individual components into larger masses, each mass with its own global properties, of which pitch is only one. It is consistent with our earlier discussion to call each such mass a stream; by doing so we point out that such global organizations have a time dimension to them as well as a spectral structure.

To summarize our discussion thus far, it appears that many factors in complex spectra affect the ability of the auditory system to parse them and uncover the individual acoustic sources that gave rise to them. These include the density of the spectra (how closely partials are spaced), the relative intensities of the partials, the match in the perceived intensity of the partials to the intensities of earlier sounds, as well as the harmonic relations between the partials.

Common Fate (AM and FM)

The factors that have just been discussed—the harmonicity principle, for example—make use of properties of a sound that remain steady

for at least a short time. The principles that we will examine next concern changes in the sound over time. Changes can provide us with a powerful rule for deciding whether parts of a spectrum belong together. The rule is this: If different parts of the spectrum change in the same way at the same time, they probably belong to the same environmental sound.

By way of introduction, I would like to describe a principle of grouping put forward by the Gestalt psychologists, the principle of "common fate." Let us imagine that we had a photograph taken of the sky. It shows a large number of birds in flight. Because they are all facing the same direction and seem to be at the same distance from the camera, we think that they are a single flock. Later we are shown a motion picture taken of that same scene. On looking at this view, we see that there were two distinct flocks rather than only one. This conclusion becomes evident when we look at the paths of motion. One group of birds seems to be moving in a set of parallel paths describing a curve in the sky. Another group is also moving in a set of parallel paths but this path is different from that of the first group. The common motion within each subgroup binds that group together perceptually and, at the same time, segregates it from the other group. The common motion within each group is an example of the Gestalt principle of common fate.

The segregating effects of common fate need not be restricted to motion. Suppose we were faced with an array of identical disks, each at a different randomly chosen level of illumination. A still photograph of the disks might cause us to group clusters of them on the basis of similarities in brightness. Because of the random brightnesses, this grouping would not yield clusters with simple shapes. However, a motion picture of that same array might reveal that the disks on the right half of the display were undergoing parallel changes (possibly only a slight flickering) in brightness, and this would cause us to see them as a group. The synchronous change would not necessarily cause the disks on the right to be more similar in brightness to one another than to those on the left. At any given moment, a still photograph would always find a haphazard array of brightnesses. The change itself is the important thing.

The common fate heuristic in vision groups any subset of elements of a scene whose changes are proportional and synchronous and it segregates them from other elements of the scene that are changing in a different way. We can see that this is a good way to do things. It is very unlikely that unrelated elements in a scene subset will undergo parallel changes by accident.

The principle of common fate also has an application in audition. Suppose it was found that two frequency components were changing synchronously by proportional amounts. This would seem to be very unlikely by chance. It is much more reasonable to assume that the two are parts of the same sound, that is, that they have arisen from the same physical disturbance in the environment. It is likely in our world that those frequency components that arise from a single acoustic source will go on and off at more or less the same time, will glide up and down in frequency together, will swell and decline in intensity together, and so on.

There are two types of synchronous changes that have been studied, changes in frequency (called frequency modulation or FM) and changes in amplitude (called amplitude modulation or AM). One would expect both of these to be useful in the listener's task of parsing the set of simultaneous components into streams.

FM: Parallel Changes in the Frequency of Partials

The first type of correlated change that we will examine is frequency modulation. An everyday example of this is in the human voice. As we tighten our vocal cords, the pitch of the voice rises. Physically, the fundamental frequency rises and all the harmonics rise in a proportional way as they must.[331] As the fundamental doubles in frequency, say from 100 to 200 Hz, the fifth harmonic must also double, from 500 to 1,000 Hz, since it always remains in the fixed proportion 5:1 to the fundamental. In fact all the harmonics must double in frequency to maintain their fixed ratio relation to the fundamental. Therefore all the harmonics must change by the same proportion. This is the same as saying that they must change by an equal amount in log frequency. If we look at a set of harmonics of a 100-Hz fundamental as it rises in frequency, the harmonics rise together on a log frequency scale as in figure 3.7. The first thing we notice is that their motion is parallel on the logarithmic scale. Second, they appear more and more crowded together as we increase the harmonic number. This latter fact arises from the relation between the harmonic series and the logarithmic scale. The successive harmonics in a tone are not spaced from one another in equal logarithmic steps. Instead they are spaced by a fixed number of cycles in frequency, and this fixed difference appears smaller and smaller as you move up a logarithmic scale. However, the choice of a logarithmic scale is not arbitrary: The mapping of frequency onto the basilar membrane also is logarithmic. Therefore, we can take figure 3.7 as a picture of how the peaks of activity on the basilar membrane will change over time.

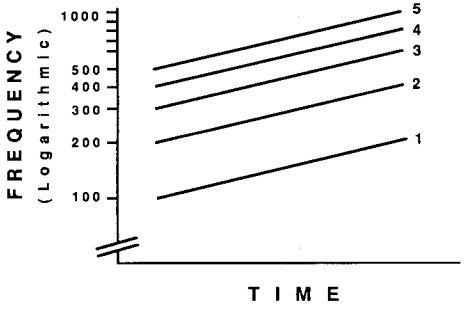


Figure 3.7
The spacing and the parallel motion of harmonics on a log frequency scale.

We would expect that since nature causes harmonics to move in parallel, our auditory systems take advantage of this regularity. In a mixture of sounds, any partials that change in frequency in an exactly synchronous way and whose temporal paths are parallel on a log-frequency scale are probably partials that have been derived from a single acoustic source. Therefore the spectral grouping process should unite them for subsequent global processes to work on. There is some evidence that the auditory system works this way.

We do not know what the physiological mechanism for this is. The fact that as harmonics move up and down in parallel they maintain a constant separation on the basilar membrane implies that there is a simple change on the receiving organ as well as in the external world. However, we do not know whether the auditory system exploits this anatomical simplicity. Before looking at the evidence for a grouping principle based on parallel FM, we must bear in mind the close relation between such a principle and the one based on harmonicity that we have already discussed. If we see the auditory system grouping harmonics that move in parallel frequency paths together, how do we know whether they have been grouped because they were *moving* together? Perhaps it was just that they always retained a harmonic relation to one another. Perhaps the fact that they were moving had no significance. Before we could attribute the grouping to the movement itself, we would have to see it adding some extra strength to the tendency for harmonics to group or else showing some ability to cause inharmonic partials to be perceptually fused. Much of the research that has been done on the perceptual grouping of modulated harmonics does not distinguish between the retention of harmonicity over time and the effects of the common motion.

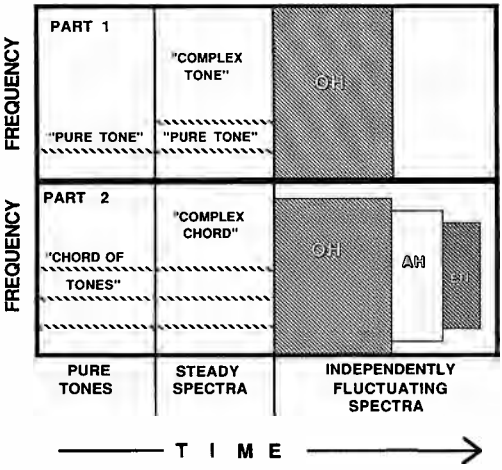


Figure 3.8
A demonstration of the power of coherent frequency fluctuation to integrate the harmonics of singing voices. Part 1: progression from a pure tone to a complex tone to the perception of a singing voice. Part 2: progression from a chord of pure tones to three singing voices. (After Chowning 1980.)

Micromodulation

In 1979, John Chowning of the Stanford University Center for Computer Research in Music and Acoustics reported some results that he had obtained when trying to synthesize a human singing voice. It appeared that the sound would not fuse and become a unitary percept unless a small amount of frequency fluctuation was added to the fundamental, and hence to all its harmonics.[332]

Basing his work on the assumption that it was the parallel frequency fluctuation that caused the harmonics to fuse, he created a fascinating demonstration that is illustrated in figure 3.8. In part 1 of the figure, we see first a pure tone at 400 Hz. It is then joined by harmonics that would be present in a vowel, in the intensities that would be present in that vowel, but with no frequency fluctuations. What the listener will hear at this point is a continuation of the 400-Hz tone, but now joined by a second tone that is harmonically complex, like an organ tone. The combination is not yet heard as a voice. In the final period all the harmonics have a coherent frequency fluctuation imposed on them; now the fusion of the set of harmonics is much better, and the sound is heard as a voice singing a vowel (for example, "oh") at a pitch of 400 Hz.

In part 2, the same game is played with a mixture of three singing

voices, one singing an "oh" with a fundamental of 400, one singing an "ah" at 500, and the third an "eh" at 600 Hz. The demonstration begins with three pure tones, the fundamentals of the three voices, at 400, 500, and 600 Hz. This is heard as a chord containing three pitches. Then the full set of harmonics for all three vowels is added, but without frequency fluctuations. This is not heard as a mixture of voices but as a complex sound in which the quality of a chord is present but the three pitches are not clear. Finally, in the third section, the three sets of harmonics are differentiated from one another by their patterns of fluctuation. We then hear three vocal sounds being sung at three pitches (300, 400, and 500 Hz).

Let me explain how the three sets of harmonics were caused to segregate from one another and yield three voices. It is possible to impose a frequency fluctuation on a fundamental and all its harmonics so that they all move together in parallel on log frequency coordinates. Chowning chose a pattern of frequency variation that was like the vibrato of a singer. It had a repeating up and down sinusoidal variation but some irregularity was also added to the pitch change so that it did not sound mechanical. Each of the three sets of harmonics had a different pattern of fluctuation in frequency, the differences being in the rates of vibrato (4.5, 5.0, and 5.5 Hz) as well as in the patterns of irregularity. As a result, the harmonics of the same synthesized voice moved in synchrony and in parallel, retaining their harmonic relations to one another over time. However, the partials from one voice had a constantly changing relation to the partials of each of the other voices. The size of the frequency fluctuation that Chowning imposed on the frequencies was about plus or minus 5 percent (or 10 percent total excursion), a typical figure for a vibrato in a singing voice.

Small fluctuations in frequency occur naturally in the human voice and in musical instruments. The fluctuations are often not very large, ranging from less than 1 percent for a clarinet tone to about 1 percent for a voice trying to hold a steady pitch, with larger excursions of as much as 20 percent for the vibrato of a singer. Even the smaller amounts of frequency fluctuation can have potent effects on the perceptual grouping of the component harmonics. Small frequency modulations of this type have been called "micromodulation" by Jean-Claude Risset, a French composer of computer music.

Micromodulation has been the object of intense study in a Ph.D. thesis by Stephen McAdams.[333] One of the important questions addressed by this thesis was whether it was necessary to have parallel motion on log frequency coordinates before fusion was induced. What about another type of parallel modulation in which the frequency

components move up and down together but instead of maintaining their frequency *ratios* as they move, maintain their frequency *differences* instead? In the latter type of modulation, if you started with three harmonics, of 100, 200, and 300 Hz, and shifted the 100-Hz component up by 37 Hz, all the harmonics would go up by 37 Hz, yielding frequencies of 137, 237, and 337 Hz. Notice that these new partials would no longer be harmonically related. This kind of change is called constant-difference (CD) modulation. It is contrasted with the more normal case where all the partials are *multiplied* by a constant amount and thus maintain their ratio relations to one another; this is called constant-ratio (CR) modulation.

McAdams started with 16 harmonics of a 220-Hz fundamental, and imposed a frequency fluctuation of about 6 per second on them all.[334] In different conditions, he imposed different degrees of fluctuations, ranging from less than 1 percent (the variation of pitch that occurs in a clarinet) to just over 3 percent (a relatively small vibrato in a singing voice). The listeners were presented with two sounds, one containing constant-ratio modulation and the other containing constant-difference modulation. They were asked to judge which one seemed to have more "sources" in it. Their overwhelming choice was that the constant-difference modulation gave the impression of more sources, their choice of the CD tone increasing from about 60 percent with the small fluctuations to 90 percent or more when the large fluctuations were used. Evidently the constant-difference modulation, instead of helping the tone to fuse, actually caused its partials to segregate from one another. For example, the fundamental frequency was clearly heard as a separate sound. The reason for the breakup should be obvious. This sort of modulation destroys the harmonic relations among the partials. Evidently the benefit of concurrent modulation of this type, if it exists at all, is outweighed by the harm caused by reducing the harmonicity of the sound.

This is another example of where the Gestalt principles are best understood within a scene-analysis framework. There is no reason to reject CD modulation as an example of the Gestalt principle of common fate. After all, it consists of synchronous and similar changes. However, CD modulation is unnatural. There is no such regularity in nature and therefore we have no reason to expect our auditory system to be able to exploit it for purposes of scene analysis.

McAdams also ran a series of experiments comparing stimuli that either had incoherent or coherent micromodulation. There were a number of experiments, but they all had the same form. A complex sound containing many partials was played. There were two basic conditions: In the first, called coherent modulation, the same jitter

pattern was applied to all the partials so that they changed in synchrony and in parallel on log frequency coordinates. In the second condition, incoherent modulation, the total set of partials was divided into two subsets by the experimenter. One jitter pattern was applied to one of these subsets of partials and a second, independent, jitter pattern was applied to the second subset. The frequency fluctuation of each subset of partials was coherent internally but was not coherent with the fluctuation of the other subset. In various experiments, he tried out frequency fluctuations as low as .01 percent and as high as 5 percent. The question was whether the ear would organize the two subsets into separate sounds.

In one experiment, for example, a single partial was given one pattern of fluctuation and all the rest received a second pattern. In different conditions within this experiment, the selected partial ranged from the first to the sixteenth. McAdams reported that in a pretest that used a 3–5 percent frequency fluctuation, the pitch of the selected partial could be heard as a separate pitch for any selected partial up to the sixteenth.[335] In the experiment proper (using much smaller fluctuations), once the size of the fluctuations reached 0.5 percent, the subjects virtually always chose the incoherently modulated sound as having more sources in it than the coherently modulated sound. The "more sources" judgment is a tricky one, however. McAdams reported that this judgment was based on different experiences depending on which partial was involved. In the case of the lower ones, the listener heard the partial standing out as a separate pitch. But with higher harmonics, the listener heard a kind of choral effect, the effect that you hear when you listen to a choir, in which you cannot distinguish individual voices but you know from some incoherence in the sound that more than one voice is involved.

These results supported the same conclusion as a pretest that McAdams reported in his thesis.[336] It was possible for trained listeners in a laboratory situation to hear out partial pitches for any of the first five to seven individual harmonics of a complex tone if the tone was sustained and unmodulated. However, when all the harmonics were modulated coherently, the listeners could no longer do this. The only pitch that they heard was the global pitch of the fundamental. Apparently in this case, as well, the coherent modulation increased the fusion of the harmonics.

McAdams also reported another pretest that showed how important coherent modulation is in encouraging the auditory system to compute the global pitch of a sound in preference to the partial pitches of its frequency components. He mixed three signals together and asked some musically trained listeners to tell him how many pitches

they heard. Each of the three was a synthesized harmonic sound with the spectrum of a vowel. When two or three of the sounds were steady (not modulated), the listeners reported sometimes hearing four to six pitches. However, when each was given its own pattern of frequency fluctuation, the listeners reported that the three global pitches belonging to the three vowel spectra were clearer and less equivocal. He argued that these results were in agreement with findings reported by other researchers that the global pitches of sounds that have little or no energy at the fundamental are stronger when the partials are modulated coherently than when there is no modulation.[337]

It appears that even the global pitch associated with inharmonic complexes gets stronger with micromodulation. Once, when I was at IRCAM (Institut de Recherche et Coordination Acoustique/Musique), the world-renowned center for computer music in Paris, McAdams played me a demonstration in which micromodulation was applied to a mixture of three inharmonic sounds. Each sound was made by stretching the harmonic series in the way that I described earlier. When these were played without modulation, I could hear a weak sense of global pitch but certainly not three separate ones; I also heard a number of partial pitches. When modulation began (a separate pattern for each set of partials), I heard the three global pitches get stronger. The only difference between what is heard with a mixture of inharmonic spectra and a mixture of harmonic spectra is that in the case of the harmonic spectra the partial pitches often completely disappear whereas in the case of the inharmonic spectra many of these partials remain audible.

McAdams has reported an interesting musical effect that was created for a composition by Roger Reynolds at IRCAM.[338] The sound of an oboe was analyzed and then resynthesized with the even and odd subsets of harmonics sent to two different loudspeakers on different sides of the room. When the sound began, the two subsets were given exactly the same pattern of frequency fluctuation and this served to hold them together as a single sound heard somewhere between the loudspeakers. However, as the sound continued, the frequency fluctuations of the two subsets gradually became independent, with a separate pattern of fluctuation for each loudspeaker. This caused the listener to segregate the two subsets and hear two sounds, one coming from each speaker. The effect was that of "a soprano-like sound an octave higher (the even harmonics) and one of a hollow, almost clarinet-like sound at the original pitch (the odd harmonics)."[339] Why the two different pitches? Because of any

tone that contains the even harmonics of a tone at frequency f (2f, 4f, 6f, . . .) can just as easily be viewed as containing all the harmonics of a tone at frequency 2f, a tone an octave higher.

This demonstration does two things. It displays the power of modulation to overcome even the classical clues to spatial location and it also makes us wonder about the order in which the qualities of a sound are computed in the auditory system. The demonstration suggests that frequency modulation affects perceptual grouping and then grouping determines perceived pitch and perceived location. In short, grouping determines location. Yet we know from other cases that cues for location can determine grouping: It is easier to follow one voice despite the presence of another if the two are at different locations. Which way does it go then? Does grouping determine location or does location determine grouping? We see this sort of pattern again and again in this book. My resolution of the contradiction is to imagine a highly interactive scene-analysis process in which all the major forms of evidence enter into the analysis. The function of this analysis is to converge on the interpretation of the scene that best satisfies the evidence.

Glides

We have seen that small fluctuations in pitch can serve to segregate subsets of partials that have different patterns of fluctuation. We would expect similar results in cases in which subsets of partials underwent longer (and slower) glides in parallel on log frequency coordinates. Only a few experiments on the fusion and decomposition of gliding partials are available.

One experiment of this type was done by Lynn Halpern and myself in 1977.[340] We asked people to listen to sounds that were made up of partials that were gliding in straight lines on a log frequency scale. Examples are shown in figure 3.9. There were always two subsets, each of which included one or more partials. One subset glided up and the other glided down in frequency. The glides were 1 second in duration. Within each subset, the partials maintained a fixed harmonic relation to one another throughout the glide. These ratios are marked on the figure. For example in part E the ascending glides maintain the fixed relation 3:4:5 throughout; that is, they could be considered to be the third, fourth, and fifth harmonics of a low tone. Similarly, in part E, the descending glides are in the ratios 10:11:12. Rather than hearing a confused mass of sound, the listeners heard two distinct sounds, one gliding up in pitch and the other gliding down. The two sounds were distinct enough that the listeners could make

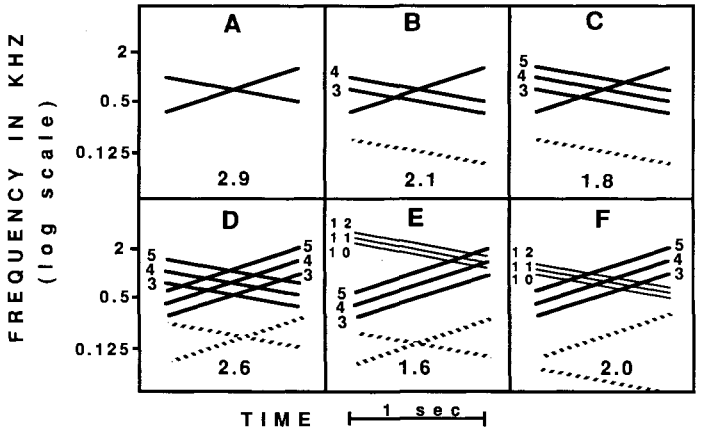


Figure 3.9
Patterns of crossing parallel glides. (From Halpern 1977.)

reliable judgments about whether the two sounds crossed in the middle or bounced apart. When they consisted of only single partials (part A), they tended to be heard as drawing together for the first half second and then bouncing apart. If we look at the figure, we can see that in the case of two pure-tone glides, the pattern is somewhat ambiguous to the eye. Nothing tells it whether the glides actually cross or meet at the middle and then bounce apart. If the eye has any preference at all, it prefers to see the lines cross. The situation is quite different for the ear. Rather than following a glide through its intersection the ear prefers to hear it stay in the same frequency range.

This preference applies whenever the two subsets of partials have the same internal frequency ratios. For example, in part D the partials in both the ascending and descending subsets maintain the ratio 3:4:5 to one another throughout their glides. Again the listener prefers to hear them approach and then bounce apart. The numbers at the bottom of each panel represent the preference (on an arbitrary scale) for hearing the tones bounce apart as measured in two replications of the experiment. Scores above 2.5 indicate that bouncing was more audible and scores below 2.5 mean that crossing was. The patterns shown in parts A and D show the strongest preference for bouncing. However, in some of the other conditions, where the within-subset ratios differ for the two subsets, the preferred organization is to hear the two sounds cross in the middle. In these cases the listener hears two tones with different qualities and uses the quality of the tone that is being focused on as a way of following it through the crossover

point. What all this means is that if there is a difference in quality between the rising and falling subsets, the glides will be heard to cross. For example, the descending set in part F sounds like the yowling of a cat while the ascending set sounds more melodious. But notice what this description implies. If the difference in quality between the subsets is to affect the preference for hearing crossing or bouncing, the quality has to be derived from within each subset. The auditory system has to have grouped together only the partials that were gliding in parallel in order to have come up with this separate assessment of quality. Parallel modulation, then, appears to encourage the fusion of partials.

Unfortunately, there is some uncertainty about the cause of this segregation. It is possible that it was due to the parallel frequency change (frequency modulation) of the glides. This would make it an example of the Gestalt principle of common fate. However, there is another equally good explanation: It was not the parallel frequency change itself that was responsible for the segregation but the mere fact that the partials, by moving in parallel, maintained their harmonic relations to one another. Since we know from our earlier discussion that harmonically related partials will fuse, we might argue that there is nothing new here. The subsets would have segregated just as well had they not been gliding, as long as the two subsets did not bear a harmonic relation to one another. As with the case of micromodulation, it must be shown either that gliding harmonic tones fuse together even better than steady ones do or that gliding can fuse even nonharmonic tones.

A more analytic experiment was done sometime later at McGill by Peter Doehring and myself.[341] I have described it earlier in this chapter when discussing the effects of harmonicity upon fusion. The stimulus pattern was shown in figure 3.6. A pure-tone glide repeatedly alternated with a set of three simultaneous equal-duration glides, one of which, the target, was always the same as the captor in all respects. Therefore, it could be captured into a sequential stream by the captor and heard as a separate tone in the mixture. The two glides that were presented simultaneously with the captor tone were referred to as the frame. The two glides of the frame were always the same as one another in slope (on a time by log frequency scale) and were always two octaves apart throughout their excursions. This good harmonic relation (ratio of 4:1) tended to encourage their fusion.

The slope of the target relative to the frame was varied in the experiment. As the slope difference increased, the glide became easier to capture. This suggests that a parallel change in frequency caused the

target to fuse with the frame. However, the results suggested otherwise. The conditions that showed this effect were all ones in which the target was positioned exactly between the glides of the frame and maintained a one octave separation from both of them. However, in other conditions the target was kept parallel to them but shifted up or down by a quarter octave from the perfect octave separation. In these conditions the target was no easier to capture when parallel to the frame than if it had an altogether different slope. The common slope alone was not enough. It seemed that a good harmonic relation had to be maintained.

Was it simply that there was no effect of the modulation and all that was important was maintaining good harmonic relations? This hypothesis fails to explain one fact about the data. The effects of the parallelness of the target to the frame were not all or nothing. A difference in slope of .5 octave per second had a less segregating effect than a difference of 1 octave per second. Yet in both cases, the deviation of slopes would effectively destroy any harmonic relations. Maybe some auditory analysis actually looks at the slopes after all. The question is still open.

It is hard to know what to conclude from this single study. McAdams maintains that parallel micromodulation will cause even inharmonic partials to fuse at least a little better than they would without modulation. In such a case there is no harmonicity to be maintained by the parallel modulation. Two explanations come to mind to explain the contradiction between this claim and the parallel-glides experiment. The first possibility is that although slow continued parallel modulation cannot fuse inharmonic partials, micromodulation can. This would make micromodulation and slower modulation different sorts of beasts. The other explanation relates to the fact that in the parallel glides experiment that Doehring and I did, the frame defined a strong harmonic framework that could have made the mistuned glide stand out very clearly. It may be that this provoked an active rejection of the nonfitting component and made it unlikely that even parallel modulation could cause it to fuse with the other glides. If all the partials were badly related to one another, as in the inharmonic stimuli that McAdams used, this positive rejection process might not occur since no strong harmonic framework would exist and therefore nothing could stand out as being mistuned from it. This hypothesis again brings up the need that I mentioned earlier for experiments to find out whether the strength with which a set of partials establishes a harmonic framework can affect the capacity of the ear to reject tones that do not fit.

AM: Common Amplitude Change at Different Spectral Locations

Onset-Offset Asynchrony

Earlier I presented an experiment that I did with Steven Pinker in which the tendency of two pure tones to fuse depended on the synchrony of their onsets.[342] There are really two ways of explaining this result. One is by the "old-plus-new" heuristic that was the subject of our earlier discussion. Briefly, this principle says "if part of a present sound can be interpreted as being a continuation of an earlier sound, then it should be." An explanation of the experimental finding by this principle would say that when two sounds come on asynchronously, one of them is first heard without the other and we can get a better fix on its properties so as to extract it from the mixture. Is this the whole explanation of the effects of the synchrony or asynchrony of the onset of partials? I think not. Although this factor surely plays a role, I believe that the synchronies themselves also play a role. This idea is attractive when you consider that in the normal acoustic world, it is exceedingly improbable that unrelated sounds will just happen to go on or off at exactly the same time. Therefore synchrony is an excellent indication that acoustic components must have arisen out of the same sonic event.

The first piece of evidence that suggests that there is an independent effect of synchrony comes from an experiment that looked at offset synchronies as well as synchronies at the onset.[343] It looked at the tendency for a group of three simultaneous pure tones (actually the harmonics f, 2f, and 4f) to fuse and resist the capturing effect of a preceding pure tone. The onsets and offsets of the three harmonics could be varied independently of one another. The experiment found that a harmonic that extended past the other two, either by coming on sooner or going off later, was easier to capture out of the mixture.

The old-plus-new heuristic was able to explain how we can segregate a partial that comes on ahead of the others. But what about the case where a later offset of a partial helps us to segregate it from the others. We also get to hear it alone in this case, but only after the mixture has gone by. Can a piece of knowledge acquired later help us to decompose a mixture that happened earlier? Or more generally, are there retroactive effects in auditory perception? There is ample evidence that there are. I am not referring to the higher-order examples in speech understanding in which, for example, the interpretation of the word "check" in the sentence "A check is not a checkmate" depends on the later word "checkmate". Such examples seem a bit more sophisticated than the simple hearing of two sounds rather than one. However, there is evidence for retrospective effects in even

simple perceptual phenomena. For example when we perceive a sound as continuing through a very noisy masker even when it is not there, this illusion depends, in part, on the sound appearing again at the moment the noise ceases.[344] This means that something heard during the noise depends on what is presented to us after the noise has ceased. We think that we heard it at the time that it happened, but we must have revised our memory of it at a later time.

Here is a puzzle: When a component comes on a little before the rest of the spectrum of which it is a part, or ends shortly after it, do we use that information to really hear that component out of the mixture or is this "hearing out" just an illusion? We might just be basing our judgment of the tone on the part that sticks out and might be getting no information whatever about the part that is inside the mixture. There is some evidence from masking experiments done by Rasch that this description is true.[345] It will be discussed later when we look at masking studies. However, there is an observation that can easily be made by anybody with several audio oscillators that will demonstrate the segregative effects of onsets and offset that do not extend the target tone outside the mixture. If we tune the oscillators to a harmonic series, we will hear a fairly thick, single sound. However, if we take the oscillator that is playing the third harmonic and turn it on and off repeatedly while the remainder stays constant, we will hear the third harmonic as a separate pulsing tone. The independent temporal changes in that harmonic will isolate it. If all the harmonics had been pulsed together, we would simply have heard the global sound pulsing.

This effect was studied carefully by Michael Kubovy.[346] He created a monaural complex of nonharmonically related partials (whose frequencies were selected from a diatonic scale) with frequency separations ranging from two to six semitones between successive partials, and played it with equal intensity for each partial. This sounded like a complex chord. However, the sound could be changed by selecting one of the partials as a target every 300 msec or so and briefly lowering its intensity and then restoring it to its original level. When this was done, the pitch of that partial became prominent. By selecting a different partial every 300 msec, he could create the experience of a tune. The emphasis that the selected partial received was not just due to the fact that it had received a boost in amplitude at a moment when the other partials had not. In a variation of the experiment, every 300 msec or so a selected partial was attenuated by 12 dB for 77 msec, then restored to its original level by increasing it by 12 dB. At the same time as the target was incremented, all the other partials were also incremented by 12 dB and remained briefly at this higher level.

Despite the fact that all the partials rose in intensity at exactly the same moment, it was the pitch of the target partial that became prominent. If successively higher partials were chosen in turn to be the target partial, the listener could hear an ascending series of pitches. It was not the rise in intensity that distinguished the target partial from the others in this case, but the order in which the rise and fall in intensity took place, the order being down-up for the target and up-down for the others. The auditory system seems therefore to have a great sensitivity to differences in the amplitude *pattern* for different partials and can use these differences to segregate them.

Another experiment done in Kubovy's laboratory showed that if one partial in a steady complex tone was suddenly thrown out of phase from the others its pitch became audible.[347] The stimulus was a tone whose fundamental was 200 Hz. It contained all harmonics from 3 to 14. They were aligned in sine phase; that is, at the instants (200 times per second) when the absent fundamental would have been passing through 0 phase angle, all the harmonics that were present were passing through 0. At certain moments in time, all the harmonics except one were instantly reset to zero phase and, at the same instant, the remaining harmonic was reset to some other phase angle. The difference in treatment made the pitch of this particular harmonic audible. A succession of interruptions of this type occurred, each choosing a different harmonic to be the odd man out. The harmonics selected for emphasis in this way formed an ascending scale on some trials and a descending scale on others, and the listeners were required to decide whether the scale ascended or descended. Their sensitivity to the phase change depended on how far out of phase the selected harmonics were thrown. When they were 20° out, the listeners were getting about three-quarters correct, and at 40° they were almost always correct.

This result seems to imply that the ear can instantaneously compare the phase of all the components and find subsets that are incongruent, but the authors of this experiment pointed out a mechanism, known to exist in the peripheral auditory system, that could change a phase difference into an amplitude difference.[348] Therefore the more central auditory system may actually be detecting the same sort of change as when the amplitude of a partial is shifted suddenly with respect to that of other partials.

So far it sounds as if different patterns of amplitude change can segregate different partials of a complex sound. But what about the reverse: Can a synchronous change in amplitude cause sounds that would otherwise not fuse to do so? Elisabeth Cohen studied our perception of a tone that was formed of "stretched partials."[349] The

frequencies of the partials of such a tone are derived from the frequencies of the harmonics of a normal tone as follows: Plot the frequencies of the harmonics on a log scale and then stretch the separations out uniformly and read off the new frequencies from the log scale. The new frequencies are no longer harmonically related. If the stretch factor is large enough, the tone will be less perceptually coherent than normal tones and we will tend to hear some of the partials as separate pitches. However, there is a method that we can use to encourage the listener to fuse these partials into a coherent tone. That is to play all the partials with a synchronous onset and an exponential decay. This percussive amplitude envelope is similar to those found in plucked instruments such as the guitar or harpsichord or in struck instruments such as the piano or gong. The exactly synchronized onset seems to tell the auditory system to fuse sounds that it would not fuse without this cue.

The role of synchrony and asynchrony of onset in the partitioning of mixtures of sounds faces us with a puzzle. In the tones of many musical instruments some of the partials rise and fall in amplitude at different times than others. For example, one study found that in a D4 (D in the fourth octave) played by a trumpet, the fifth, sixth and seventh harmonics reached their maximum amplitudes about 20 msec later than the first three harmonics did.[350] Why do we not hear the delayed harmonics as defining a separate sound? There may be a number of reasons.[351] First, the the asynchrony is small, usually less than 20 msec for most instruments. Second, there is often a noisy onset of the tone, which serves to mask the asynchrony of the onsets. In addition, the rise times of the various partials may not form two distinct clusters, and therefore distinct onsets may not be detected; if there is simply a mush of partials coming on at different times, the onset of any one of them may not be distinct enough from the others to cause the partial to be treated as a separate sound. In other cases, the later-rising partials may increase too gradually in intensity to be detected as a separate tone. Finally, since we know that fusion competes with sequential streaming, there may not be an ongoing stream at the frequency of the delayed partial to capture it, and it may remain in the stream of its own instrument. On the other hand, even if there is a capturing stream (perhaps a partial from some other instrument in a complex orchestral piece) the listener may simply group the asynchronous partial with the spectrum of the other instrument without noticing that it is gone from the spectrum of its own instrument. In short, under conditions of isolation the decomposition is not likely to happen, and in a mixture it is not likely to be noticed.

There is further confirmation of the role of asynchrony in segregating components of sounds. This evidence comes from experiments on masking. It is harder to mask a target if it comes on when the masker is already present.[352] However, we will discuss this evidence later when we consider the findings of masking experiments and how they relate to questions about the simultaneous organization of sound.

The partitioning of spectra by means of synchronous amplitude changes is consistent with our understanding of neurology. A neural model for the grouping of parts of the auditory spectrum has been proposed by Malsberg and Schneider.[353] The synchronous activation of neurons that are responding to different parts of the spectrum causes them to remain in synchronized oscillation for a short period of time. By this means the spectrum is broken into regions that these authors call segments. Neurons are synchronized within each segment and desynchronized between segments. Later recognition processes can get an unobstructed view of one of the segments by oscillating in synchrony with it.

Common Periodicity (AM) at Different Spectral Locations

As we saw earlier, there is good reason to believe that the basilar membrane of the cochlea decomposes the spectrum of the incoming sound into what we can think of as a neural spectrogram. If we look at a spectrogram of a complex sound, speech being a good example, we find that there are different temporal patterns going on in different temporal regions. As Schubert and Nixon put it,

> . . . in our immediate classification of the sounds of continuous speech . . . in the recognition of fine temporal nuance in musical performance, and particularly in our ability to separate simultaneously present broad-band sound sources, such as the instruments of an ensemble or competing talkers, there is convincing evidence that the system must either include an analyzer for direct coding of the original broad-band waveform or must routinely coordinate internally derived temporal patterns from different spectral locations in the cochlea.[354]

The question that we are examining is whether something about the temporal patterns themselves, as they occur in different spectral regions, allows the listener to put them together.

Even earlier, in 1957, Donald Broadbent and Peter Ladefoged had showed a clear understanding of this issue.[355] They, too, put it in terms of activity on the basilar membrane. How, they asked, could we understand speech when other simultaneous speech was present?

The regions on the basilar membrane that responded to the spectral bands containing energy from the two sounds would be intermixed. To segregate them, they argued, "the neural messages from sense organs [in those regions of the basilar membrane that are] stimulated by different formants of the same voice must resemble one another in some respect; and differ from the messages leaving sense organs stimulated by irrelevant sounds." They proposed that the distinguishing mark might be that the neural output of every region of the basilar membrane that was stimulated by the same sound would have a matching periodicity, since all these outputs would pulse at the fundamental frequency of the signal. We will discuss the research that was stimulated by these ideas later in a section on speech perception. For the moment it is sufficient to note that their thinking implied that common amplitude modulation in different spectral regions (and hence a common rate of neural periodicity) could bind regions of the spectrum together into a single perceived sound.

It was pointed out by others that

> Many real-life auditory stimuli have intensity peaks and valleys as a function of time in which intensity trajectories are highly correlated across frequency. This is true of speech, of interfering noise such as 'cafeteria' noise, and of many other kinds of environmental stimuli. We suggest that for such stimuli the auditory system uses across-frequency analysis of temporal modulation patterns to help register and differentiate between acoustical sources.[356]

In other words, when we are listening to a signal from a single source, the intensity changes at different places in the received spectrum tend to be correlated and the auditory system should be able to make use of this fact.

The analysis of correlations in firing rate at different parts of the basilar membrane may be an incidental by-product of the temporal mechanism, that I described earlier, for finding the fundamental frequency of a complex tone by analyzing the pattern of timing in the neural impulses activated by the sound.

Here is an example of how such a mechanism might work in detecting a correlation in the temporal fine structure at different frequency regions. When we utter a vowel sound, the acoustic result is a set of harmonics spaced from one another by F0 (the fundamental frequency). The resulting neural output from the the basilar membrane depends on the filtering activity of that organ. The output has been modeled by Richard Lyon, who displays the results in what he calls a cochleagram.[357] The latter is meant to represent the nervous

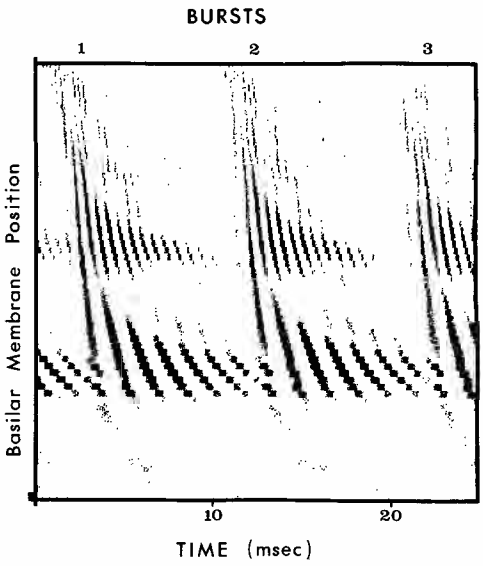


Figure 3.10
A cochleagram of a brief speech utterance, produced by Richard Lyon's computational model of cochlear output; see Lyon 1982.

system's version of the acoustician's spectrogram. An example of a cochleagram is displayed in figure 3.10. Its x axis represents time and its y axis, frequency. Darkness represents the strength of neural output at each time and frequency. We can see large temporal fluctuations in darkness (labeled bursts 1, 2, and 3) that represent beats in each frequency channel. The finer vertical striations (curving off to the right in the lower frequencies) represent the neural following of the individual frequencies. Recall that the beat rate is dominated by the fundamental. The lower fine striations are at simple multiples of the fundamental. All this regularity is extremely convenient. It means that if the auditory system groups those neural outputs from the basilar membrane where related periodicities are detected, it will probably be grouping the harmonics of a single voice, since it is extremely unlikely that two voices will be at the same fundamental frequency at the same time. If two or more voices are actually at the same fundamental frequency at the same time, as when a number of voices are singing in unison, we know that we experience great difficulty separating the individual voices from each other. To use the beats that occur at different regions of the spectrum, the auditory system would have to have a mechanism that grouped those regions

of the spectrum that were exhibiting the same amplitude modulation at the same time, and would have to be sensitive to modulations that were at the rate that corresponds to the fundamental of the human voice (roughly, from 80 to 300 Hz). We shall see below that there is evidence for the existence of such a mechanism.

The alert reader will have come up with an objection by this time. I have proposed that the auditory system could decide which regions of the spectrum of a voice have come from the same source by examining the beat rate of local regions of the spectrum. The objection is that we already know of a grouping mechanism that could do the same job without any concern for the beats. This is the mechanism that groups harmonically related partials. It simply has to decide that sets of partials in different frequency regions are harmonically related to the same fundamental (a task it has to do anyway to calculate the pitch). Why should we postulate another redundant mechanism? In the interests of parsimony we should not, for it seems that whatever cases of grouping the beat-detection mechanism might explain could easily be explained by the harmonicity mechanism.

This argument is plausible except for two facts. The first is that the human sensory systems—because they are capable of processing many features in parallel—are probably not as concerned with parsimony as theorists are. They may, indeed, have evolved redundant mechanisms because no single mechanism (or subset of mechanisms) is 100 percent reliable. The second is the fact that we can imagine situations (such as when the sound is not harmonic) in which the beat detection mechanism might solve problems that the harmonicity mechanism might not. An example is the regular pulsing of the spectrum of a motorboat, where the pulsing appears in all spectral regions but the sound is not strongly harmonic. Another occurs in the human voice in the sound "z", which is a mixture of a noisy sound with a voiced sound. In the "z", the higher-frequency noise component, made by the front of the tongue, is actually being powered by the glottal vibration, which is periodic. This imparts the periodicity of the fundamental to the noise. So although the noise component is not harmonic, and could not be grouped with the voiced component by a harmonicity analyzer, it contains the same periodicity of amplitude fluctuation as the voiced component. Therefore a mechanism that tended to group different spectral regions by similarity of amplitude modulation could put together the sound components of the "z" as easily as it put together the parts of the motorboat sound. To summarize, we can say that the harmonic analyzer cannot deal with noisy sound, but an analyzer for beats or local periodicities could deal with both inharmonic and harmonic sounds with equal facility.[358]

Before looking at the evidence for whether similar periodicities at different spectral locations could be used to group portions of the spectrum for the purposes of subsequent pattern recognition, we should ask whether there is any evidence to suggest that the auditory system has the capacity to compare, or to combine, the periodicity information in different parts of the spectrum. One piece of evidence that it can do so comes from a study of pitch perception.

We must first recall that the pitch of the fundamental of a set of harmonics is often perceived when the fundamental itself is not present. The most convincing explanation of this is one which sees the auditory system as registering the firing rate of neurons in each separate frequency region. After this is done, the system compares all the firing rates and calculates a fundamental rate of which the rest are all multiples.[359] The theory that firing-rate evidence is accumulated is supported by the finding that the pitch seems to be detected better when more harmonics are present.[360] There is independent evidence that periodicities that are too low to hear as pitches, but are heard as beats, can also be combined across the spectrum to yield the perception of a beat rate of which all of the detected ones are multiples.[361] Although this accumulation of information about periodicities in the spectrum in different frequency regions is concerned with the detection of a fundamental firing rate, this apparent ability makes it more plausible that the auditory system could use a common periodicity for a second purpose: to group those regions that were related to one another. The only added capability that would be needed to achieve the latter result would be the ability to detect more than one fundamental at a time and then to put the spectral regions that were related to different fundamentals into their own separate groups.

The detection of different periodicities may be responsible for some part of the segregation of ordinary tones of different pitches. Magda Halikia, in a Ph.D. thesis done at McGill, studied the perceptual segregation of two complex tones that were mixed and presented binaurally over headphones.[362] Each tone was composed of up to 32 different harmonics (different numbers in different experiments). It lasted for 1 second including gradual onsets and offsets of 200 msec. The two tones were either at the same pitch or were separated by different numbers of semitones. The fundamentals of the two were always placed symmetrically above and below 140 Hz on a log frequency scale. The listeners were asked to write down how many sounds they heard, one or two. When the two had the same frequency, they said (of course) that they heard only one tone (because mixing two identical tones simply gives the same tone at a higher intensity). But as the fundamentals of the tones were moved apart in fre-

quency, each increase in separation brought a greater tendency to judge that there were two. Interestingly enough, the greatest increase in segregation came with the change between no difference and a difference of 0.5 semitone, where the mean number of tones reported by the group of listeners changed from 1 to over 1.8. That is most of the time, at 0.5 semitone separation, the listeners heard two tones. This separation was equal to about 4 Hz. The sensitivity to the difference in fundamental frequency might actually have been more acute, but 0.5 semitone was the smallest separation used.

Halikia wanted to know whether the auditory system had to be able to resolve the frequency of the individual harmonics in order to segregate the two tones. Therefore she included conditions in which she gradually simplified the tones by removing their lower harmonics. But even when the two fundamentals were defined only by harmonics above 2,600 Hz (that is, above the eighteenth harmonic) the listeners almost always judged that there were two tones when their (missing) fundamentals had only a .5 semitone separation. This is puzzling because it is generally agreed that the adjacent harmonics in the region of the eighteenth harmonic of a 140-Hz tone could not be distinguished by any mechanism that looked for different sites of maximum stimulation on the basilar membrane. Furthermore it is generally believed that the perception of pitch cannot be stimulated by such high harmonics.[363] Nevertheless, these upper harmonics were sufficient to evoke a judgment of twoness when they were multiples of two different fundamentals. It is likely that the physiological mechanism responsible for the segregation was one that detected two different periodicities in the spectrum. Unfortunately, Halikia did not report the subjective experiences of her subjects. It is possible that they did not actually hear two tones when there were only very small pitch separations and very high harmonics. In such signals, small separations would give rise to prominent audible beats at a large number of multiples of the difference between the fundamental frequencies. It might have been the presence of the beats, rather than a clear perception of two tones, that prompted the listeners to judge that more than one tone was present.

Amplitude Modulation of Subsets of Partials We have been examining examples that suggest that the detection of matching periodicities at different sites along the basilar membrane could help to partition the spectrum. We have seen how these similar firing patterns could occur because of a common fundamental. Yet the grouping of spectral regions by shared periodicity may not be restricted to this one case. In his book *Introduction to the Psychology of Hearing*, Brian Moore de-

scribes a demonstration that shows how common changes in a set of frequency regions can cause those regions to group with one another perceptually and to segregate from others that are not changing in the same pattern:

> For example, if we present a steady complex sound containing many components with randomly distributed frequencies, a single noise-like sound will be heard, with a certain timbre. A subgroup of components in the complex can now be made to stand out perceptually from the rest by making them vary in a coherent way in either frequency, amplitude, or both. This group will be perceived as a prominent "figure" against a steady background, and both the figure and the background will have a timbre different from that of the original unvarying sound.[364]

A series of experiments have been done in my laboratory at McGill to study the integration of different spectral regions that showed the same amplitude modulation. These experiments followed the pattern of the Bregman-Pinker experiment in which a simple sound, A, is alternated with a mixture of two simple sounds, B and C, in such a way that A can capture B into a sequential stream (AB AB AB . . .). The new twist in the experiments was that A, B, and C were all pure tones that were amplitude modulated with a raised cosine wave (a sinusoidal pattern that starts at zero and oscillates between zero and one) at a frequency of about 100 Hz. That is, the amplitude of each tone rose and fell in a sinusoidal pattern about 100 times per second. The purpose of the experiments was to show that if tones B and C were both modulated at the same frequency, they would show a stronger tendency to fuse, so that B would not be free to group with A into a sequential stream.

When a sine wave is modulated with a sinusoidal pattern there are two equivalent ways of describing the result. These are shown in figure 3.11. Suppose, for example that tone B is a 1,500-Hz sine tone modulated at 100 Hz. We can describe it as I just did: as a 1,500-Hz sine tone rising and falling in intensity 100 times per second. This is shown in the upper half of box 1 of the figure. However, if we look at its long-term spectrum we will discover that there are three frequency components in it, not one. Not only is there a component at 1,500 Hz, but ones at 1,400 and 1,600 Hz as well, each having half the amplitude of the 1,500-Hz tone. This spectral result of the amplitude modulation (AM) is shown in box 2. The two additional frequencies introduced by AM are called "sidebands" and the modulated tone is called the "carrier" (these names originated with their functions in AM radio transmission). The sidebands lie above and below the

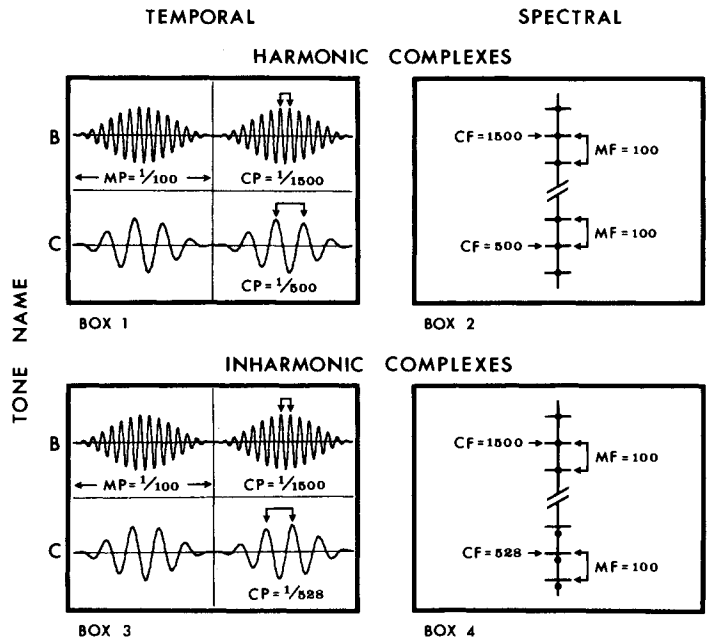


Figure 3.11
Stimuli used by Bregman, Abramson, Doehring, and Darwin (1985). Each of boxes 1 to 4 shows a representation of tones B and C. Boxes on the right, showing frequency components as the horizontal lines crossing the vertical scale, are spectral representations of the waveforms shown in the corresponding boxes on the left. MP, modulation period; CP, carrier period; MF, modulation frequency; CF, carrier frequency.

carrier frequency, separated from it by the modulating frequency (MF). Since in this example, the modulating frequency is 100 Hz, the sidebands are 100 Hz above and below the 1,500-Hz carrier.

Suppose that tone C, the one that we wanted to fuse with B, was at 500 Hz and was also modulated at 100 Hz. This would create sidebands for C at 400 and 600 Hz. Again the three resulting partials would be spaced by 100 Hz, as we see in the figure. If we look at tones B and C together, we realize that the modulation has created a situation in which all the partials are actually harmonics of 100 Hz. If such modulation were actually found to cause B and C to fuse, we would not know whether it was because they rose and fell in amplitude together, as shown in box 1, or because they simply fell into the same harmonic series, as shown in box 2. We already know that the harmonicity relation will promote fusion, so we would not know whether there was also an effect of the amplitude modulation itself.

The fact that the three components were all harmonics of a 100-Hz fundamental occurred because the carriers that I chose for this example, 1,500 and 500 Hz, were harmonics (exact multiples) of 100 Hz. Had I chosen a modulating frequency of 105 Hz, the three components that came from the 1,500-Hz carrier, 1,500, 1,605, and 1,395, would not be harmonics of any audible frequency. (Of course they are all multiples of 5 Hz, but this frequency is too low for the auditory system to hear as a missing fundamental pitch. Therefore it treats the partials as an inharmonic set.) Selecting the right modulating frequency enables an experimenter to choose whether the resulting components will or will not be related to a common audible fundamental. This means that it is possible to arrange it so that the modulation of two carriers by the same modulator does not automatically place them into the same harmonic series. There is also another way to ensure that the partials of C do not fit into the same harmonic series as those of B. We can choose a carrier frequency for C, (an example might be 528 Hz) that is not a harmonic of the 100-Hz modulating frequency that is being applied to both B and C. This situation is shown in boxes 3 and 4 of the figure. The AM has the same periodicity (100 Hz), but the partials in C are no longer harmonics of a fundamental of 100 Hz.

There is another problem that we have to guard against as well. If we change the harmonic series of C by changing its modulating frequency (keeping its carrier at 1,500 Hz), we create an inharmonic set of partials. One might think that now that it was not harmonically related to any audible fundamental it would have no pitch at all. However, this is not true. Apparently the auditory system computes the best-fitting pitch. Raising the modulator raises the pitch, but by much less than the raise in the modulator. This pitch, although weaker than a pitch based on true harmonics, creates a difficulty for our experiment. Although we have done away with harmonicity we have not done away with pitch. It could be argued that if we cause two carriers to fuse by modulating them at the same frequency, they will end up with the same pitch, and this, not common amplitude fluctuation, will cause them to fuse. Fortunately there is a way around this problem, because it is possible to create sounds with the same pitch by using different combinations of carrier frequency and modulating frequency. Optionally, we can keep modulating frequency constant while varying the pitch.

All these factors were taken into account at once in an experiment done in our laboratory.[365] Three complex tones, A, B, and C, were presented in a cycle with tone A first, then a mixture of B and C.

Table 3.1
The eight stimulus conditions used by Bregman, Abramson, Doehring, and Darwin (1985).

	Pitch	
Periodicity	100	105
100	(1) Harmonics of 100 no shift HARMONIC (400, 500, 600)	(2) Harmonics of 100 shifted up (+28 Hz) INHARMONIC (428, 528, 628)
105	(3) Harmonics of 105 shifted down (−28 Hz) INHARMONIC (392, 497, 602)	(4) Harmonics of 105 no shift HARMONIC (420, 525, 630)
	Pitch	
Periodicity	100	95
100	(5) Harmonics of 100 no shift HARMONIC (400, 500, 600)	(6) Harmonics of 100 shifted down (−28 Hz) INHARMONIC (372, 472, 572)
95	(7) Harmonics of 95 shifted up (+24 Hz) INHARMONIC (404, 499, 594)	(8) Harmonics of 95 no shift HARMONIC (380, 475, 570)

Each cycle, including the silences between the tones, lasted for 750 msec. Tone A was always identical to tone B and tended to capture it out of the BC mixture. Both A and B were always generated by a 1,500-Hz carrier modulated at 100 Hz (by a raised cosine). Therefore they were always harmonic and had a pitch related to a 100-Hz fundamental. The experiment consisted of varying the structure of tone C and seeing how this affected its ability to fuse with B and to prevent the latter from being captured out of the mixture. Table 3.1 shows the eight versions of tone C that were used. In each cell of the table, the number of the condition (from 1 to 8) is shown in parentheses. Each version of C was created by modulating a tone in the region of 500 Hz by a sinusoidal modulator (raised cosine) of about 100 Hz. The headings on the left show the exact frequency of the modulation. This is referred to as the periodicity of the resultant complex.

Let us focus on the pitch and periodicity of tone B, which was always the same. It had a periodicity of 100 bursts per second and a pitch related to a 100-Hz tone. Furthermore it was a harmonic tone; that is, it was formed of harmonics of 100 Hz. The side headings of

the table show the periodicity of the C tones. All the tones shown in the first and third rows of the table (conditions 1, 2, 5, and 6) had the same periodicity, their modulating frequency being 100 Hz. The remaining C tones had a periodicity of 105 Hz. The headings across the top show the pitch of the tones. This pitch was determined by listening to the inharmonic tones and matching them to either a 100- or 105-Hz harmonic tone (in the upper half of the table). The lower half of the table is like the upper half, but instead of some of the periodicities being at 105 Hz, they are at 95 Hz.

Some of the complexes (three-component tones) are marked as harmonic complexes and others as inharmonic in the table. We can think of the inharmonic complexes as having been made by shifting the frequencies of the three partials of a harmonic complex up or down by 28 Hz. This was found, by trial and error, to change their apparent pitch by about 5 Hz and shift it to the desired value. However, it did not change their periodicity (burst rate) because this is related to the frequency separation of the three components and we have not changed this. The actual values of the three partials (carrier and sidebands) are shown in each cell.

The results are shown in figure 3.12. In it we see the results of two slightly different experiments (the solid and dotted lines) but the results are very similar. The bottom x axis is labeled as frequency separation. This factor is important because some of the conditions brought the harmonics of C closer to those of B and this caused B and C to fuse more strongly. The y axis represents the degree to which the C tone segregated from the B tone. We can see first that all the lines slope upward to the right. This means that moving the partials of C further from those of B increased the segregation of B from C. The different slopes of the solid and dotted lines simply show that this frequency-separation effect was stronger in one experiment than in the other.

The top x axis has a different labeling. It shows what the pitch of C was and whether it matched that of B. If the pitch match made a difference, the conditions under the "100" label should be stronger than the ones under either the "105" or "95" labels, since tone B had a pitch of "100." There was no sign of any effect of pitch match. Finally the two lower lines in the graph represent cases in which the periodicity of B and C was matched and the two upper lines relate to the conditions in which periodicity was unmatched. Obviously we would not see such a separation on the graph unless the mismatch in periodicity tended to segregate B from C. The difference in segregation between tone B and C in the conditions in which their periodicities were mismatched (3, 4, 7, and 8), as compared to those in which

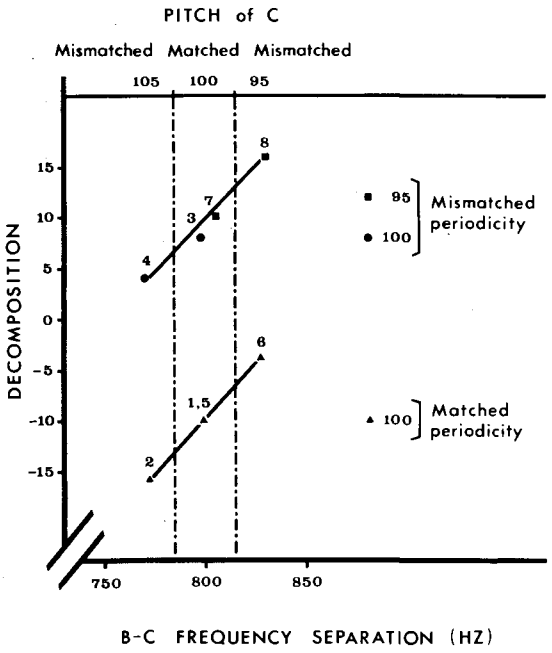


Figure 3.12
Results from experiment 1 of Bregman, Abramson, Doehring, and Darwin (1985).

their periodicities were matched (1, 2, 5, and 6), was the strongest difference found in this experiment. Furthermore, the results from the inharmonic conditions (2, 3, 6, and 7) did not deviate in any way from this overall pattern. We are able to conclude that periodicity itself has an effect on segregation independently of the effects of harmonicity, pitch, and the frequency separation of C from B.

A word of warning: The differences look very large on the graph, but that is only because a very sensitive method was used. Just listening naively to the stimuli, it does not sound as though the match in periodicity between B and C has a very strong effect on perceptual segregation. Probably the smallness of this effect was due to the use of such a narrow range of periodicities (100 Hz versus 105 or 95 Hz). But this had to be done to manipulate the pitches of the inharmonic C complexes appropriately.

Another reason for the small effect of AM may have been the fact that it was sinusoidal. Earlier, we saw that the natural-life relevance of grouping of spectral regions by their amplitude modulation was the fact that if we pass human speech sounds through a series of

band-pass filters, as we believe that the auditory system must do, we discover that the amplitude in each channel is modulated by the fundamental frequency of the voice. But unlike the case in the experiment of Bregman et al., this is not a sinusoidal modulation. The effect of the modulation, in each pitch period of the voice, is to cause a sudden rise in the intensity of the carrier signal followed by a gradual dying out. Perhaps the auditory system can use this natural form of modulation more easily than the sinusoidal modulation used in this experiment. However, the experiment used the sinusoidal form because its effect on the spectrum of the resulting complex tone is simple and its effect on the pitch of the complex is fairly well known. Modulating with a fast-rise, slow-decay waveform would spread the spectrum of the complex tone out more, so that the neural effects of modulation of different carriers would be more overlapped on the basilar membrane. We were hoping to examine the perceptual integration of signals that were neurally separated to begin with.

There is, perhaps, a way of using the more natural modulation without spreading its effects across a large number of neural channels. This would be to study the effects of AM on the integration of neural channels in patients who are fitted with multichannel electrical stimulators implanted in their cochleas. To the extent that the electrical signal sent to different locations in the cochlea actually stimulated only a restricted local region, it might be possible to study the effects of different patterns of modulation on the integration of the neural information coming from different parts of the cochlea.

A related experiment was done, again using an alternation between a single tone, A, and a mixture of two other tones, B and C.[366] Once more A, B, and C were pure tones that were amplitude modulated. As before, A and B were always identical and remained unchanged throughout the experiment. Only the structure of C was varied. Two things were varied: the match between the modulation frequencies applied to tones B and C and whether the partials of B and C did or did not fall into the same harmonic series. The listener's task was to rate the clarity of hearing tone A repeating again in the BC mixture that followed it. Tone B was always a 3,000-Hz carrier amplitude that was amplitude-modulated at 125 Hz.

Figure 3.13 shows the results for three different conditions, using different carrier frequencies for tone C: 2,000 Hz, 2,050 Hz, and 1,951 Hz. The different modulation frequencies that were applied to the carrier are shown on the x axis. The y axis shows the fusion between B and C, the lower the value the stronger the fusion. The results are quite clear in showing that the closer the modulation frequency of

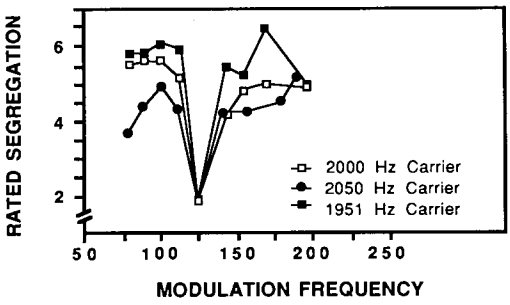


Figure 3.13
Segregation of tone B with changes in the AM match between tones B and C. (From Bregman and Levitan 1983.)

tone C approached the 125-Hz modulation frequency of tone B, the more strongly B and C fused.

There is a second aspect worth noting in the figure—the effects of harmonicity. Since the modulation frequency of tone B was always 125 Hz and its carrier frequency was 3,000 Hz (a multiple of 125 Hz), all partials of B were harmonics of 125 Hz. This was not always true for C. Recall that three different carrier frequencies were used for tone C in different conditions. One of these carriers, 2,000 Hz, had some unique properties. When it was modulated at exactly 125 Hz, it, too, gave rise to three harmonics of 125 Hz. Furthermore, when it was modulated at any of the other modulation frequencies that were used, it always gave rise to a trio of harmonics of that modulation frequency. (It was evenly divisible by all the modulation frequencies.) The other two carriers did not have this property. When modulated at any of the modulation frequencies, they gave rise, instead, to an inharmonic trio of partials.

However, the harmonicity of C made little difference. The figure shows that for all three carriers, when they were modulated at the 125-Hz frequency that matched tone B, fusion of B and C occurred regardless of harmonicity. Perhaps there is a slight tendency for the harmonic condition (2,000-Hz carrier) to fuse better, but you have to squint to see it. As in the previous experiment, the effect of a match in amplitude modulation did not depend on harmonicity.

There is another point to mention. When the C carriers were modulated at a frequency different than 125 Hz, it did not matter whether this caused C to be harmonic (within itself) or not. The curve for the 2,000-Hz carrier is not higher than the others at the AM frequencies that did not match B's. Therefore although there may be a hint of a tendency for the harmonicity of B with C to fuse them, there is no

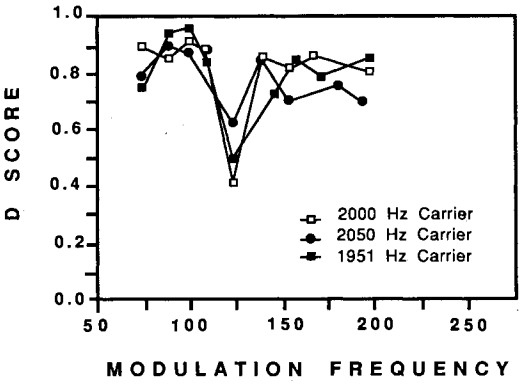


Figure 3.14
Results of accuracy judgments in an experiment on fusion by AM. (From Bregman and Levitan 1983.)

evidence that their segregation is increased when C belongs to a different harmonic series (as opposed to belonging to no harmonic series at all).

The experiment that we have just discussed depended on a subjective judgment by the listener—how clearly tone A could be heard repeating again in the BC mixture (recall that tones A and B were identical). We might wonder about the validity of accepting a listener's rating of a subjective experience. This concern motivated a second experiment done with exactly the same stimuli but with a different task.[367] On some trials, tone B was present as before but on others it was deleted. The listeners had to guess whether it was there or not and were scored according to whether they were right or wrong. The reasoning behind the experiment was this: If B and C truly tended to be more segregated in some conditions than in others, the listener's judgments of the presence or absence of B should be more accurate in these segregated conditions. As in the previous experiment, the carrier frequency for tone B was fixed at 3,000 Hz and its modulation frequency at 125 Hz. Two ranges of carrier frequency were used for tone C in different conditions. In some, the carrier was in the vicinity of 2,000 Hz as before; in others, it was in the vicinity of 1,600 Hz. Sometimes the trio of partials in C was harmonically related (carriers at 2,000 Hz and 1,600 Hz) and sometimes not.

The results, shown in figure 3.14, displayed certain similarities to the ones found with the clarity-judgment task, but also some differences. First the similarities. We see that when C had the same mod-

ulation frequency as B, the presence or absence of B was harder to detect (low scores). The difference from the previous results is that this effect of AM was much stronger when C was formed from carriers of 2,000 Hz and 1,600 Hz. These were the carriers that generated harmonics of 125 Hz (the fundamental of tone B) when modulated at 125 Hz (the modulation rate of tone B). That is, they matched B in terms of both periodicity and harmonic series. Therefore the fact that B and C were composed of harmonics of the same fundamental made B particularly hard to detect.

Harmonicity had a much stronger effect in this experiment, where the task was merely to detect B's presence, than in the previous one, where it was to rate how clearly B could be heard. What this could mean is that when harmonicity is violated, the presence of an inharmonic B causes a dissonance to arise that colors the mixture so that it can be discriminated from the case in which there is no B present; this is an additional effect that is not the same as really being able to hear B in the mixture. Although I believe that there is enough evidence from other experiments that harmonicity and inharmonicity promote fusion and segregation, respectively, we have to be sure that in any particular case, the apparent effect is not the result of the listener's being able to accomplish the experimental task without actually being able to hear the target as a separate sound. The experiment that I just described can be called an accuracy task, and such tasks are often thought to be more objective than rating tasks, where the listeners rate how clearly they can hear the target. However, they may purchase this objectivity at the expense of measuring something quite different from the perceptual effect that the experimenter is interested in.

So far we have looked at the role of the match in periodicity of amplitude pulsing at different spectral locations. Even a 5-Hz discrepancy in these rates of pulsing can reduce the tendency of these regions to fuse in perception. This requires the auditory system to have a fine level of sensitivity. However there is another experiment that shows that the system has an even finer sensitivity. It can distinguish between two pulsation patterns of the same frequency but of different phase.[368] The experiment resembles those we have just been discussing. Each of tones A, B, and C, was created by sinusoidally modulating a carrier frequency, and again tone A alternated with a mixture of B and C. However, tones A and B were not identical. Tone B was always either higher or lower than A in frequency and the listener had to report whether the high or the low version of B was heard. In all cases, B and C were modulated at the same frequency. Furthermore, B and C were both harmonics of the 100-Hz modulating frequency and therefore their partials were all harmonics

of this frequency. All this should cause them to fuse together strongly. What was varied was the phase of the amplitude modulation applied to B and C. In one condition it was in phase; the amplitudes of both A and B rose and fell in synchrony 100 times per second. In the other condition, again the amplitudes of A and B rose and fell 100 times per second, but 180° out of phase (when A rose B fell and vice versa). The results showed that there was a greater fusion between tones B and C in the in-phase condition. The listener was less able to decide whether B was higher or lower than A. This sort of result suggests that the auditory grouping mechanism may not be looking at the frequency of modulation itself, but merely at whether increases of intensity (for example) are happening synchronously at different spectral locations.

Although phase information (or, perhaps, synchrony of changes in amplitude) may affect the grouping of spectral information for some purposes, it does not seem to affect all forms of spectral grouping. For example, there is evidence that the grouping of information from different frequency regions or bands *for the purpose of deriving a pitch* does not require a phase match of the intensity variations in different frequency regions. We know that a periodic waveform (say at 200 Hz) composed of harmonics with random phases and amplitudes will give a pitch that is appropriate for the repetition rate of the waveform and that this pitch will remain even when the fundamental and the next few harmonics are removed.[369] This means that the pitch is not derived by finding points in time where the pulses in all frequency regions are in the same phase. There would be no such points (because of the random phase). Instead one would have to be able to compute the actual repetition rate of the neural activity within each frequency region, and somehow compare this computed rate in different frequency regions rather than comparing the raw pulsations themselves.

How can we reconcile the following two facts: (1) that pitch computation (really a computation of the repetition frequency of the waveform) seems to be a within-region computation that does not care about the relative phases in different frequency regions, and (2) that the grouping of regions in the experiment that we just examined does depend on phase? One possibility relates to the size of the effect. Grouping may care about phase, but not very much. If separate computations of within-band periodicity are compatible across bands, grouping may be favored even when the phases are out of line. However the alignment of phases might improve it. To investigate this possibility we would have to compare the strength of the pitch percept when the phases in different frequency bands were either

aligned or random. A second possibility is that the grouping of pitch information across frequency regions does not follow the same rules as the use of periodicities to group the regions for the extraction of other properties. It is too early to decide between these alternatives.

Perhaps the following experiment could tell us how important the synchrony of pulses in different regions was, independently of local pitch estimates. Suppose we used a series of short bursts of a sinusoid, where the carrier had a fixed frequency but where the bursts were scheduled to occur at random intervals varying, for example, between 2 and 20 msec, with a mean burst rate of 100 per second. We could have this occur with carriers of two different frequencies, with either the same schedule of bursts governing the two burst trains or else two independent schedules with the same overall statistical properties. In this case, if the synchronies themselves were important, the two frequency bands should fuse much better when they were both activated with the same schedule for the bursts. One would expect stronger differences between the synchronized and unsynchronized conditions than were obtained with the experiment on out-of-phase AM that I described earlier. In that experiment, the bursts in one frequency region always had a fixed delay relative to the bursts in the other (one-half of the modulation period). The regularity of this relation may have reduced the segregating effects of the asynchrony of the bursts. The auditory system may treat an irregular asynchrony as offering better evidence for the existence of separate sources than a fixed and repeating asynchrony does. A mere displacement of the pulses in one spectral region relative to those in another is not really very good evidence that two separate sounds are occurring. A fixed-delay asynchrony will occur, for example, between the amplitude changes in the low frequencies of a sound and the high frequencies of its echo (which may be mixed with the original sound). A statistical independence in the changes is a much better clue that there are two sounds than is a mere asynchrony of the changes.

Simultaneous amplitude changes help us to appropriately fuse or segregate sound that comes not only from different parts of the spectrum but also from different places in space. If you present different tones to the two ears, for example, 750 and 800 Hz, they will not fuse. But if you add an in-phase amplitude modulation of 8 Hz to both, they will.[370] An interesting side effect is that the single sound image seems to be localized on the side of the ear receiving the higher frequency. (This localization toward the ear receiving the higher tone also occurs in Diana Deutsch's octave illusion.[371])

We might ask ourselves whether a tendency to fuse the information from different directions when their pulsation is synchronized has any

value in a real-world listening situation. The clue to its utility might be the fact that when we are standing close to a wall or other reflective surface, a significant amount of energy may be reaching one of our ears by a path that distorts the information about its true spatial origin. Furthermore, because the wall may be absorptive, the energy coming off it may have a different spectrum than the direct energy does. If the amplitude pulses are slow enough that their phase misalignment (resulting from the fact that they have traveled by different paths) is not large in comparison with the period of the pulsation, their approximate synchrony may provide information that the different spectra received at the two ears are really parts of the same source spectrum. It might be that the 8-Hz modulation demonstration gives us a glimpse of a mechanism evolved to exploit this sort of cue.

I have left out some studies that show strong effects of correlated AM in different frequency bands in uniting the information from those bands. They are concerned with masking, and I want to discuss all the masking experiments together later in this chapter.

We have seen that even though two sources of acoustic energy (for example, two instruments) are sounding at the same time and their partials are mixing at our ears, their momentary changes in intensity may tell us that there are two groups of partials.

We have examined the auditory system's use of correlated changes in amplitude in different spectral bands as evidence that the bands were created by the same physical source. We have only looked, so far, at rapid fluctuations in intensity. However, we must also consider slow changes. These might occur, for example, because of the spatial movement of either the source of sound or the listener. The partials of an approaching source will all become more intense together at our ear relative to the partials from sources that are not approaching. The same thing will happen when we ourselves are moving toward some sound sources and away from others or even when we turn our heads.

The issue of *changes* in intensity is unfortunately complicated by the issue of momentary *differences* in the intensity. If one physical source (a set of partials, A) swells and becomes considerably louder than another simultaneously active source (B), the spectral shape of the mixture will undoubtedly be very close to that of the spectrum of the louder sound, A, taken alone. The memory for that spectrum, created at moments when A is much louder, may make it possible to isolate A's spectrum from the mixture of A and B at a later moment when A is not emphasized by loudness.

When sources are changing in loudness relative to one another, therefore, we do not know whether to explain our ability to sub-

divide a mixture by appealing to a mechanism capable of using the correlations of the changes in loudness or one capable of exploiting momentary differences in loudness. We could imagine mechanisms that worked in either way. For example, a mechanism that looked for correlated intensity changes would first measure the intensity change over some brief time period for each resolvable part of the spectrum and then look for regions of the spectrum that showed the same changes. We have already discussed how useful such a mechanism might be in segregating a human voice from other periodic sounds.

However, a mechanism that used its memory for earlier events would also be useful, particularly in isolating a part of the spectrum in a very dense mixture. In such a mixture, especially where the components of two or more sounds were overlapped in frequency, the overlap might prevent the listener from obtaining separate assessments of the local properties of the different regions of the spectrum (such properties as frequency or amplitude changes or direction of arrival) so as to correlate them with properties found elsewhere in the spectrum. In such a case, getting clear "peeks" at the individual spectra would become particularly useful. Fortunately, sounds are rarely steady. The human voice, for example, changes constantly in frequency composition and in amplitude. If two voices are mixed together, there will be moments when one is relatively quiet and we will get a good peek at such properties as the spatial location and the shape of the spectrum of the other one. Even those moments when the components of one spectral region in one of the voices cease may allow us to get a peek at the components of the other voice in that region. To take advantage of these moments, the auditory system would have to be able to keep a record of them and then compare them with, and possibly group them with, later-arriving parts of the mixture.

Although it is hard to imagine exactly how this could be done by the auditory system, we already know that something very much like it would have to be done in order to carry out the old-plus-new heuristic that we discussed earlier. Recall that this heuristic examines a mixture of components to see whether it contains components that were present alone just before the mixture occurred. In order to carry this out, the auditory system would need a record of the components that were present at that earlier moment. In the experiments of the type done by Pinker and myself, memory for a single simple sound would be adequate. However, in real-life environments, the system could never know when a new source would become active and add its spectrum to the mixture. Therefore, its record of the components that were candidates for extraction from a later complex spectrum

would have to be continuously changing and represent the entire recent spectrum.

I think there is every reason to assume that it is the same sequential grouping process that forms the sequential streams discussed earlier that records the peeks we are discussing now. If we make this assumption, it can suggest properties that a peek-using process might have. For example, we know that when we rapidly alternate two tones of quite different frequencies, the stream segregation that occurs can group a sound not just with the previous sound but with an earlier one. Therefore we suspect that the peek-using process has the ability to keep a record not only of the most recent sound but of earlier ones as well. How this could be done is a mystery, but our attempt to unify a variety of grouping phenomena strongly suggests that it must happen. Perhaps the record that is kept is more like a spectrogram (a time varying "picture") than like "the most recent spectrum," and the process that uses it can match up "subpictures" that are separated in time but resemble one another.

The strong likelihood that a listener has the capacity of recording time-varying spectra and following them into subsequent mixtures creates a problem for us as theoreticians: When we want to assert that the auditory system has decomposed a mixture by making use of correlated amplitude *changes*, we must try to rule out the use of simple momentary amplitude *differences*, which can give the system a peek at one of the component spectra. At the present time there has been, as far as I know, no systematic investigation of these issues.

Comparison of AM and FM Effects

We have seen how the grouping of regions of the spectrum can be affected by two sorts of correlations between changes in those regions. These are correlation between variations in frequency (FM) and in amplitude (AM). I have described AM and FM as different sorts of changes in the signal, but the auditory system may not always detect them in a different way. Let me describe a case in which the auditory system may be treating FM as just another form of AM. First let us set look at the cochlear transformation of the incoming signal and then see how it might be responding to AM and FM. The cochlea may be viewed as an array of band-pass filters, each of which is centered on a different frequency and has a neural output that shows a maximum when the input frequency is at the tuned frequency, a sharply falling off of response to frequencies above the tuned frequency, and a more gentle decline of response to frequencies below it. This is shown schematically in figure 3.15. The asymmetrical shape

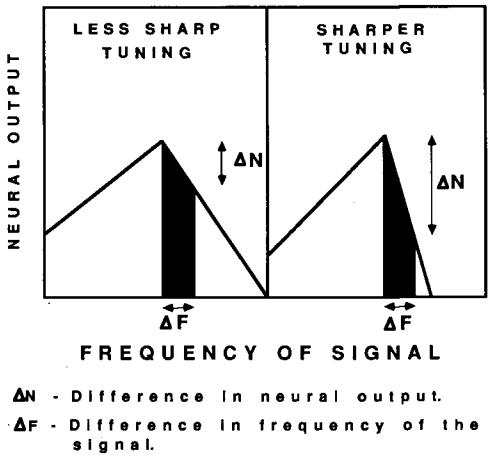


Figure 3.15
The magnitude of neural output of a local frequency region of the cochlea to inputs of different frequencies.

of the filter transfer function that is shown here summarizes the finding of a large number of neurological and psychophysical studies. Although the asymmetry is more or less agreed upon, the sharpness of the tuning is not. What is important in the figure is that the filter seems more sharply tuned on the high-frequency side, showing large changes in neural output (labeled N in the figure) for small changes of the frequency of the input (labeled F). Because of this sharp tuning, frequency changes are converted to changes in neural output.

Let us consider what happens in response to modulation in the frequency of a partial. The effects of two different assumptions about the sharpness of the filtering are shown. Suppose that the modulation moves a partial from the tuned frequency of a filter to a frequency somewhat higher. The intensity of neural response from that filter will drop substantially. When the frequency comes down again, the intensity of response will rise. As we see by comparing the two halves of the figure, the change in output activity for a fixed change in frequency depends on the sharpness of the tuning.

Now let us imagine that the partial that we are considering undergoes a changes in amplitude rather than frequency; again the neural output intensity will rise and fall in step with the modulation. We can see, then, that both FM and AM can cause the filter outputs to modulate in intensity.

It is therefore theoretically possible that the perceptual grouping of different spectral regions that are changing in a correlated way may be

based on this modulation of the bursts of neural output both for AM and FM. This does not mean that listeners cannot tell the difference between frequency changes and amplitude changes. Clearly they can. However it could still be the case that the grouping process treats them as equivalent.

Pierre Abdel Ahad and I carried out an experiment to try to throw some light on this question.[372] It compared how well different types of modulation promoted the grouping of frequency regions. The type of stimulus pattern resembled that of Bregman and Pinker, as shown in figure 1.16 of chapter 1. It was a repeating cycle composed of a tone A, alternating with a mixture of two tones, B and C. Tones A and B were identical and the listener's task was to try to hear tone A repeating in the BC mixture. As in many of our experiments, A, B, and C were not pure tones but complex ones. The characteristics of tone C were varied in different conditions to see how this affected the listener's ability to segregate B from it. Each of tones A, B, and C consisted of one or more sinusoidal components. The components within a tone were all modulated in the same way. Tone C was either modulated in phase with B or in exactly opposite phase and it was the difference between these two conditions (in phase versus out of phase) that was the focus of interest. The rate of modulation was 6 Hz, which is slow, but at a rate that is not unlike the vibrato of the human voice.

So far I have not said anything about the type of modulation that was used. There were, in fact, three types, but only the first two are relevant to the present discussion. The first was amplitude modulation (AM). This was sinusoidal in form with a frequency of 6 Hz. Amplitude modulation can be described by its depth. That is, the modulation need not change the signal from full amplitude to inaudible. In this experiment, the difference between the most and least intense phase of the signal was either 4, 8, or 16 dB. We wanted to see how different depths of AM applied to tones B and C would affect their fusion when the modulations were either in phase or out of phase.

The second type was frequency modulation (FM). Frequency modulation can be described by its excursion factor, that is, by how far the frequency of the carrier is varied around its central frequency. It is convenient to express this as a percentage of the frequency being modulated, because if we want to preserve the harmonic relations among the components being modulated, we must modulate them all by the same fixed percentage of their frequencies. In this experiment we used excursion factors of 2, 4, and 8 percent. Again these were applied to tones A and B either in phase or out of phase.

We wanted to find some level of AM that gave the same effects on the segregation and fusion of tones B and C. From this, using the ideas shown in figure 3.15, we could estimate what the slope of the cochlear filter must be if AM and FM were actually both detected by their effects on the changes in the amount of neural output from individual filters. If this estimate gave a plausible value for the slope, it would support the idea of a common mechanism for AM and FM effects.

In order to accommodate a third condition (whose description is omitted here) we made each of the tones A, B, and C out of three partials, a carrier accompanied by two sidebands that were separated from the carrier by 100 Hz and had half its amplitude. The carrier frequencies for tones B and C were always 3,600 and 2,400 Hz respectively, and tones A and B were always identical.

The 6-Hz modulation had different effects in the AM and FM conditions. In the AM condition, the intensity of the three partials (e.g., those in A) simply rose and fell as a group six times per second. In the FM condition, the frequency of each component of the three-component tone rose and fell by the same excursion factor six times per second, maintaining the same harmonic relations to one another throughout.

The task of the listener was to reduce the intensity of tone C (which started off very intense) until A and B could be heard together in their own stream. The results are shown in figure 3.16. The loudness of C that could be tolerated while still hearing a separate sequence of tones A and B is shown as a function of the amount of modulation for the two forms of modulation. Higher levels of the curves represent increased degrees of segregation of tones B and C. For both forms of modulation, the segregation is greater when the modulation of B and C is out of phase than when it is in phase. Since the results are in decibels, they may be directly compared across conditions.

Comparing the FM and AM conditions, and interpolating a little, we see that an 8 percent excursion produced by FM is worth about the same as perhaps a 10-dB change introduced by AM. Using the reasoning represented in figure 3.15, we can calculate that the results of AM and FM could be the result of their effects on the intensity of the gross bursts of neural output if the falloff of the cochlear filters had a slope of at least 90 dB per octave. This is a simplification based on the assumption that the falloff is symmetrical in the low-frequency and high-frequency direction. We know that this is not true and that the falloff in the low-frequency direction is slower. Since we know that the falloffs observed on the high-frequency side of the tuning

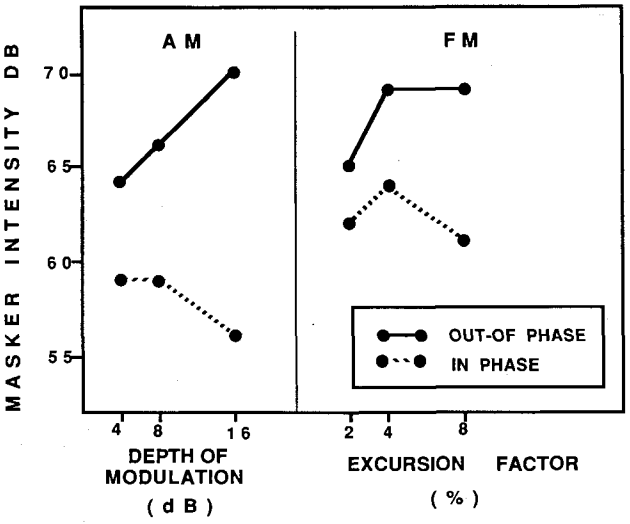


Figure 3.16
Results of a comparison of the effects of AM and FM on perceptual fusion. Left panel: AM results. Right panel: FM results. (From Bregman and Abdel Ahad 1985.)

curves of primary auditory neurons can be in the range of 200–500 dB per octave, there is more than enough sharpness in the tuning of the local regions of the auditory system to account for the grouping effects of FM on spectral fusion purely in terms of its effects on the *amplitude* of neural pulsations.[373]

A second experiment was done comparing the effects of these two forms of modulation. It was identical to the previous one except for the fact that tones A, B, and C had much denser spectra, 12 partials each. The results were much the same with these denser spectra as with those of the previous experiment.

These results suggest that it is possible that the same neural mechanism may be responsible for the effects of AM and FM on the perceptual grouping of frequency bands. They do not rule out the possibility that an entirely different mechanism underlies the two effects, but they show that a common mechanism is not implausible given the relative sizes of the two effects. If the FM effects had been much larger than could be explained by amplitude changes in basilar membrane filters, we would have had to postulate that FM was detected in a different way.

Correlation of Auditory with Visual Changes

We do not just live in a world of sound. Many of the events that make sound also have visible effects. Furthermore the effects on the sound and the light will often correspond in time. This correspondence is another example of the Gestalt concept of common fate. As I hit the keys of the keyboard upon which I am writing this chapter, there is an audible click at a certain phase of the motion. A person listening and watching should be able to utilize this temporal correspondence to determine which sounds, in a collection of intermingled sounds, are coming from the typing. If the position of the motion in the visual field is more clearly defined than the spatial origin of the sound, we could draw the inference that the sound was coming from a corresponding location in auditory space. We might then be able to direct our auditory attention to that location. Conversely, if the auditory location were clearer, the correspondence would tell us where to direct our visual attention. In other words, the correspondence, in time and in space, of auditory and visual changes should allow auditory and visual computations to help each other.

There have been a number of studies on how visual and auditory information is blended to allow the perceiver to know where in space an event is taking place, or what the identity of that event is.

One sort of study concerns the "ventriloquism" effect in which a sound appears to come from the spatial location in which a correlated visual event is occurring. This has been the subject of a great deal of study.[374] Another coupling between audition and vision that has been experimentally studied is the tendency of infants to prefer to look at objects whose motions are correlated with changes in a sound that they are hearing, or who show in other ways that they are treating the auditory and visual information as coming from the same object.[375] As I mentioned earlier, there is also evidence that the tendency to integrate a rapid sequence of auditory events into the same stream depends on whether a sequence of visual events, synchronous with the auditory ones, is seen as integrated into a single stream of motion.[376] Despite all this research, there have been no studies of whether the correlation of a subset of sounds with a sequence of visual events can assist the listener in extracting that subset from the auditory mixture.

There is perhaps one exception to this lack of research. There have been studies showing that if we see a speaker's face in a noisy environment, we can more easily understand what is being said.[377] However, there is reason to believe that in the case of speech perception the visual events are not just telling us that the acoustic components

correlated with it are coming from the same source. Instead, the visual modality is actually supplying information about what the person is saying. We know how to recognize words both auditorily and visually. When we hear a word and see it spoken at the same time, we combine the evidence from the two sources. Experiments have been done in which the information sent visually (in a video image) and auditorily (in a synchronized sound track) was discordant.[378] For example, "ba-ba" might be presented auditorily and the word "ga-ga" visually. The actually perceived sound would be one that is as consistent as possible with both sources of information. In this example, we would hear "da-da", since it agrees with the visual "ga-ga" in being spoken with the lips open and with the auditory "ba-ba" in being a voiced stop consonant that is articulated at the front of the mouth. The integration is not a conscious one. Knowing what is happening does not prevent the illusion.

However, this form of unconscious integration must surely depend on our knowledge of the language and therefore does not fall into the category of scene-analysis processes that we have referred to as primitive ones. These latter exploit regularities that occur in a wide class of listening situations, not just in speech. It is possible that the primitive, general process of integrating simultaneous sounds and visual events is involved even in the auditory-visual integration of speech. However, it is hard to prove that it is since the highly trained skill of listening to speech is functioning at the same time and perhaps obscuring the effects of the primitive process. This difficulty is particularly evident in interpreting the duplex perception of speech, discussed in chapter 7.

Although there is no formal research on the role of auditory-visual temporal correspondences in primitive scene analysis there are some informal examples that suggest that they may be used. If you want to help a person to hear something in a mixture—for example, in an orchestral piece—a strategy that is often used is to synchronize the gestures of your hand with the loudness swells or rhythmic accents in the target sound. Here you correlate the property of intensity in your hand and the sound. In common language, you beat out the notes. Another strategy is to correlate the vertical position of your hand with the pitch of the notes as you beat them out. There seems to be a natural association between the following pairs of properties that change at the same time in vision and audition: (1) changes of intensity of motion (speed and amplitude) in vision with changes in amplitude of sound, and (2) change of vertical position in space with change of pitch.

Summary of Effects of Common Fate

We have seen that the Gestalt concept of common fate has provided us with a way of talking about the auditory system's use of a number of different types of clues that could tell it whether to treat co-occurring auditory events as having arisen from the same signal. A number of clues have been mentioned. These have included corresponding changes in frequency in different parts of the spectrum. These could be the large slow changes that occur when the voice is gliding in pitch or the very tiny fast ones when there are rapid fluctuations in the voice. These correlated changes lead to diverse effects. One is an improvement of the quality of the pitch sensation derived from a subset of partials. Another is the isolation of the timbre of the set of correlated partials. Single partials that change in ways that are uncorrelated with the others are heard as pure tones and a group of partials changing together define a single sound with its own perceptual qualities. Unfortunately we cannot tell from the existing research whether all these effects can be explained by the fact that the parallel changes in log frequency that have been used in most of these studies had their effect simply by maintaining good harmonic relations within subsets of partials.

Correlated amplitude changes are also important. These changes include the synchrony or lack of synchrony of onsets and offsets, and of changes in amplitude (and possibly in phase). The detection of these amplitude changes can act in support of other detected changes, as when the beats between harmonics that are detected in auditory "filters" occur at the same rate as the fundamental and give support to evidence that might be found from a more direct assessment of the spacing of the harmonics on the basilar membrane. On the other hand they can oppose other clues, as when a rapid onset followed by exponential decay can lead the listener to fuse a set of inharmonic partials that would otherwise not have fused.

Finally, there is some suggestion that correlations between some spectral activity in audition and some visual activity will help to isolate acoustic components from a mixture.

All these clues can be described by the same principle: If there is a correspondence between a change in an auditory component and in something else (either another auditory component or a nonauditory event) this is probably not an accident and the auditory component should be assigned to the same perceived event as the other change with which it is correlated.

Spatial Correspondence

We will now shift our attention to one of the strongest scene-analysis principles. This is the one that says that acoustic components that can be localized as coming from the same position in space should be assigned to the same stream. We know that the cue of spatial location is a very important one in allowing a listener to follow the words of one speaker in a noisy environment.[379] If we cover one of our ears in such a situation our ability to select the desired voice gets much worse.

It is evident that the frequency components arising from the voices of different speakers will be interleaved in frequency, and that the early spectral analysis done by the auditory system will not in itself be sufficient to segregate them. To compute a separate identity for each naturally occurring sound in the mixture, the auditory system has to be able to group the overall set of resolved frequency components or frequency regions into the subsets that represent the individual sounds. Could this grouping be done according to the spatial origin of the components?

There is a difficulty involved. At the stage in processing that grouping must be done, the system is still working with frequency components, not whole sounds. Therefore, in order to know how to group these components, it has to be able to assign a spatial origin to each one separately. Only then could it know which ones came from the same location. What evidence is there that the auditory system can compute location on a frequency-by-frequency basis?

We saw in an earlier chapter that the principle of grouping sounds that come from a common spatial origin was used to create separate perceptual streams for sounds that were rapidly alternating between the ears. However, we are now dealing with events that happen at the same time, not in succession. We know intuitively that we can hear two sounds at the same time and know that they are coming from different spatial locations. But what more do we know about this impressive capability?

Kubovy has presented an argument that does not assign a central importance to spatial location in audition.[380] He has argued that while it may be true that space is an "indispensable attribute" in vision, it is frequency that plays that role in audition. An indispensable attribute is a feature of the sensory input that permits the perception of the twoness of two simultaneous signals that are identical except for that feature. In vision, two otherwise identical objects that are either separated in time or in space can be seen as two things. Therefore space and time are both indispensable attributes in vision.

In audition, Kubovy argues, the indispensable attributes are time and frequency. Frequency is considered indispensable since two simultaneous sounds differing only in frequency can be heard at the same time. The role of time (in the range of the longer time intervals that do not give rise to the perception of frequency) is the same as in vision. But space, according to Kubovy, is not an indispensable attribute in audition. He argues that two sounds that differ only in their spatial origin will be fused in perception and heard as emanating from some intermediate location. (We shall see shortly that this is not always the case.) Similarly loudness is not an indispensable attribute of audition, since two otherwise identical and simultaneous sounds of different intensities coming from the same position in space will simply sum at the ear and be heard as one.

Since frequency and time are Kubovy's indispensable attributes of hearing and we have already seen that they appear to be the ones involved in stream segregation, it would be tempting to guess that only his indispensable attributes are used in the segregation of simultaneous sounds. Unfortunately for this hypothesis, this volume contains counterexamples that suggest that spatial separation plays an important role. It is true that a spatial difference alone cannot cause segregation of two simultaneous tones that go on and off at precisely the same time. However, as soon as the situation becomes more complex, two simultaneous frequency components differing only in place of origin can be heard as parts of two separate sounds at different spatial locations rather that uniting to form a sound heard between those locations.[381] In addition, when a spatial difference is added to some other difference (say in frequency or in a match to a prior event in one ear), it greatly increases the amount of segregation.

Evidence That Ear Comparisons Must Be Frequency Specific

Let us return to the question of whether the spatial origins of different frequency components can be assessed independently. There is some indirect evidence from physiology. If a cat is operated upon and parts of its primary auditory cortex are lesioned, it may no longer be able to tell where a sound is coming from in space when tested with a short burst of a certain frequency; at the same time its spatial localization for other frequencies may be intact and normal.[382] Larger lesions may wipe out the ability to localize most frequencies but may not affect the ability to localize the frequencies that are handled by the part of the cortex that is still intact. Furthermore the deficit will occur only in the half of the auditory field on the opposite side to the lesion. The cortical locations that will produce deficits at different fre-

quencies are arranged in regular strips along the cortical surface. The results cannot be attributed to simple deafness to those frequencies resulting from by the operation. Frequency-specific deficits have also been reported in humans with brain damage. These results support the idea that the auditory system is capable of frequency-specific localizations.

There is also evidence gathered in psychoacoustic experiments that show that two pure tones of different frequencies at different locations can be heard simultaneously. When one is presented to each ear over headphones, they will not fuse to generate a single image unless the tones are very close in frequency. For example, when the frequency of one tone is 250 Hz, the other one must be less than 6.6 Hz (3 percent) away from it before fusion occurs. In the range from 250 to 4,000 Hz, the maximum mistuning that is tolerated is never greater than about 7 percent.[383] Therefore this fusion is quite finely tuned.

This experiment measures fusion, but not its opposite—independent localization. There will have to be extensive research on separate judgments of locations for simultaneous pure tones before we can know exactly how finely tuned the auditory system is for making separate frequency-specific localizations. The research will be hard to do because the mere simultaneity of the tones will tend to fuse them. We may need special indirect methods for the study of such localizations.

Such an indirect method was used in a fascinating experiment done by Kubovy and Howard that showed not only that closely spaced parts of the spectrum can be localized separately from one another but that the separate location information is remembered for some time.[384] They played to listeners a chord of six pure tones ranging from 392 Hz to 659 Hz. These tones spanned a range of less than an octave, and corresponded to the musical pitches G, A, B, C, D, and E. All the tones were sent to both ears over headphones, but the two ears received slightly different temporal information about each tone. It is possible, in the laboratory, to specify a location for a tone along the left-right dimension by delaying the phase of the sound a bit in one ear relative to the other. This simulates what happens in real life: the sound takes a little longer to get to the ear that is further from the source. Taking advantage of this cue for location, Kubovy and Howard introduced a different delay in phase for each tone, thus specifying a different horizontal location for each one. Even when this was done, the overall sound was confused, because no single frequency stood out of the mixture. Then, after a brief pause, the chord was played again, but the specified location of one of the tones was changed while the others were held constant. Now the tone with the

changed location popped out of the mixture and its pitch was heard faintly, separate from the confused mass. The change was made during the silence between the two presentations so that each of the presentations, if heard alone, sounded like a blurred mass. It was only when the listeners heard one presentation shortly after the other that they heard a tone pop out. The researchers argued that in order to know which tone's location had changed on the second presentation, the auditory system must have had separate records of the previous locations of all the tones. It would have had to preserve these records until after the silence and then compare them to the new ones that came after the silence. Over how long a silence could these records be preserved? For most of the people in the experiment, the silence had to be less than 1.5 seconds. However, one unusual listener could detect the changed tone with 100 percent accuracy even across the longest silence that was tested, 9.7 seconds. Therefore not only did this experiment demonstrate the existence of a computation of location that is separate for each frequency, but also a separate memory for each.

At this point I would like to cite not a scientific study but a sound pattern that was created at IRCAM, the computer music center in Paris, for use in a piece of music.[385] It was created by synthesizing the even and odd harmonics of an oboe tone on separate channels and sending the channels to two different loudspeakers, to the left and right of the listener. A small amount of frequency fluctuation (FM micromodulation) was imposed on the harmonics. At the beginning of the sound, the fluctuations of all the harmonics were synchronized and identical, and the listener heard a single oboe centered between the speakers. Then the fluctuations of the harmonics in the left speaker were gradually made independent of those on the right, but within each speaker the harmonics fluctuated in synchrony. As the two channels became independent, the sound seemed to split into two separate sounds, one based on the even and the other on the odd harmonics, coming from separate speakers. We can understand why the two sounds would be heard as one when the fluctuations were correlated. But in the case of the two uncorrelated channels, how could the listeners know which harmonics to hear in each speaker unless they were able to assign independent locations to the individual harmonics and then group those that came from the same location and had a common micromodulation? The different qualities of the two perceived sounds furnished evidence that the segregation had actually caused the listeners to group the same-channel harmonics together. The sound coming from the speaker containing the odd

harmonics had the characteristic, hollow, clarinet-like quality associated with a sound composed of only odd harmonics. Furthermore, the sound composed of the even harmonics sounded, as it should, an octave higher. The amazing thing is that the even and odd harmonics are closely interwoven in frequency and that not only the two pitches but the two qualities were correctly assigned to separate locations.

A reader might object that there is nothing unusual about hearing the sound that comes from a speaker as actually coming from it, or in hearing two sounds coming from the places that they are actually coming from. However, the initial fusion of the two sets of partials shows that this accomplishment is not inevitable, and makes us more sensitive to the magnitude of the accomplishment. We must remember that the auditory system does not have access to the two subsets of partials as whole signals. It first takes the sounds apart into their frequency components. If it is to put together all those that came from the same place, it must first figure out which place each of them came from. When the micromodulation in the two channels was correlated this acted to promote fusion; that is, it operated to contradict the directional cues, causing the partials to blend even though they did not all come from the same place.

This is a good example of how different cues will compete to control the fusion of the components. The usefulness of this particular competition in natural situations is that sometimes we may hear a sound in which a single signal has been bounced off different reflectors on two sides of us before it reaches our ears. In such a case, the reflectors may each absorb different frequency bands in different amounts, causing the input to one of our ears to have quite different acoustic parameters than the input to the other. A simple interaural intensity comparison (for example) might find different points of origin for different frequency components. If, however, the different inputs shared a common frequency-modulation pattern, this common-fate cue might cause the components to be fused, thereby overcoming the distorting effects of the environment.

There is another phenomenon that shows the effects of the ear's ability to make a separate estimate of point of spatial origin for different frequency regions. This time the segregation protects the weaker sound from being masked by the louder. The phenomenon is called the "binaural masking level difference" (BMLD) effect. I will discuss it later in a section that collects together different masking effects that are related to scene analysis.

Here is yet another demonstration of the ear's ability to separate different spectral components on the basis of their points of origin in space. Let us start with a broad-band noise that is created by a single

white noise generator and choose a narrow band of its spectrum, lying below 1,000 Hz.[386] Then, for the frequency components in this part of the spectrum only, we delay the phase of the copy of the signal going to one ear relative to the copy going to the other. For the rest of the spectrum we send identical copies to both ears with no time difference. This acoustic stimulus causes the listener to hear a pitch that is like the pitch of a narrow band of noise at the ear that gets the earlier version of the narrow-band signal. This pitch stands out faintly against the rest of the broad-band noise, which is heard as centered. The segregation does not depend on the sound actually starting in one ear first, but on the ongoing disparity in phase between the two ears. As long as the phase leads continuously at one ear, the signal is heard at that ear. The isolated pitch cannot be heard by either ear taken alone but depends on the cross-ear comparison. In other words, the band of noise in which the components are out of phase at the two ears, as they would be if an environmental sound were on one side of the listener, is assigned a location different from the other spectral components. This segregates the frequency bands and the pitch of the narrower band is heard. The ears are acting as if there were two different sounds in the environment.

An important sort of fusion takes place when the echo of a sound is fused with the original so that we hear only a single sound. This interpretation is valid since only one sonic event has actually occurred. If the fusion is a sort of scene-analysis phenomenon, we would expect it to be affected by the spatial origins of the two sounds. For example, if the original sound came from straight ahead, the strength of fusion should depend on where the echo came from. If it also came from straight ahead (echoing off a wall behind the source), it should show a greater tendency to fuse with the original than if it came from the side (bouncing off one of the side walls). This is exactly what has been observed.[387]

As a brief interruption of our discussion of spatial effects on fusion, we might consider other plausible expectations about echoes. It is reasonable to expect that any of the factors that influence the segregation of two sounds would affect the segregation of an echo from an original. For instance, if the room resonances were to color the echo, giving it spectral characteristics different than the original, we would expect that the two would segregate more readily. The research on the role of onset asynchrony in segregation (as well as common sense) would tell us to also expect the amount of asynchrony to be important.

Thinking of the suppression of perceived echo as a scene-analysis problem leads us to note the following: A sound that is varying in

frequency or intensity will be accompanied by an echo that is undergoing an identical series of changes, but with a certain delay. This common fate, even though slightly out of phase, may promote the fusion of the sounds. However, this common fate could be detected only if the ear's "covariation detector" could detect covariations even if they were displaced in time. The question of whether the auditory system can do this deserves experimental study.

The evidence that we have reviewed points to a capacity to segregate different components of a total acoustic input into streams that come from different locations. The perceptual result that we have mentioned thus far has been an ability to separate the subsets that come from different locations. But how about the ability to perceptually integrate such subsets? The following experiments suggest that, under some circumstances, we cannot put them back together even when an experimenter asks us to do so.

David Green and his associates at Harvard University explored an auditory capacity that Green called "profile analysis." Listeners can learn to discriminate two different spectrum shapes independently of their loudness. For example, if two sounds are played with different overall loudnesses and listeners are asked whether they have the same spectral shape, they can correctly judge them as different even if the difference involves a modest increase in the intensity of one spectral component (at 949 Hz) *relative to the others*. That is, boosting the intensity of the 949-Hz component in the spectrum makes it sound different and this difference is not just a loudness difference but a difference in some globally computed quality. Now let us suppose that we separate the 949-Hz component from the others and send it to the opposite ear. Can the listener now integrate the spectrum across the two ears so as to evaluate the *shape* of the overall spectrum without regard to any change in overall loudness? Experimental results have suggested that the listeners could not integrate the left and right ear sounds into a single qualitative analysis.[388] They simply heard the left and right ear signals as two individual sounds. Sometimes the segregation by spatial location is quite compelling.

The tendency to segregate sounds that come from different spatial locations can help us hear them more clearly. In San Antonio, Texas, there is a type of grackle that likes to congregate with its fellows in the same large tree, shrieking loudly. Although one of these birds is not unpleasant to listen to, a large number of them calling at slightly different times and pitches creates an unholy cacophony. Once when I was listening to this sound, I thought of covering one ear to see what the effect would be. Despite the lowering of loudness, the sense of dissonance and roughness rose. Uncovering the ear, I realized that

I was now aware of more individual components, even though the intermingling was such that the individual sounds poked their head, as it were, out of the mixture for only brief instants, giving the sound a sort of glittering quality. When one ear was covered these highlights were not as evident and the whole sound seemed more smeared and dissonant.

This effect seems to show that even the partial segregation of a multiplicity of sounds through spatial segregation prevents the auditory system from computing certain dissonances between them. As we shall see later, when we examine the case of musical dissonance, there seems to be evidence that if we can prevent the individual notes of a potentially dissonant chord from fusing by capturing these notes into separate auditory streams, the combination will not sound dissonant. Dissonance seems to be strongest when all the components are part of a single undifferentiated stream. Perhaps this lack of differentiation was what occurred when I covered one ear in the presence of the grackles.

Not all the evidence for the role of spatial differences in enhancing the segregation of signals is as impressionistic as this. In the 1950s there were a number of studies that showed that if a person were asked to repeat one of two mixed verbal messages, it helped a great deal if the two voices were presented over different headphones.[389] To cite another example, Robert Efron and his associates at the Veterans Administration Medical Center in Martinez, California, in the course of studying the effects of brain damage, presented a mixture of five sustained natural environmental sounds, such as violin, saw, cat meow, ping-pong, and human voice, to normal control listeners. They found that these were better identified when the sounds were made to seem to come from different locations (inside the head of the listener) by separately adjusting the loudness balance for each sound in the left and right headphones.[390]

Computer Programs That Use Spatial Correspondence The idea of using the direction of origin to segregate sounds has occurred to engineers designing computer systems to segregate simultaneous voices. Apparently if the number of microphones is larger than the number of voices and the location of each voice is known in advance, it is possible for a machine to filter out all but the desired voice. However, it is harder do this with just two microphones, analogous to the two human ears. In this case one cannot, according to theory, eliminate the offending voices, but it is possible to attenuate them.[391] One attempt required the target voice to originate at the center of a rectangle formed by four microphones.[392] If a different voice had been

designated as the target, the microphones would have had to be rearranged so that the new target voice was at the center of the rectangle. A computer program operating on the input from these four microphones processed the signal. It did not split the signal into separate frequency bands and assess their spatial origins separately as we believe people do. However, it had some success; when a recording of the processed signal was played to listeners, the unwanted speech was both attenuated and distorted. The distortion of the unwanted speech seemed to increase the intelligibility of the target speech. This method seems far inferior to the capabilities of persons who work with only two ears that do not have to surround the target signal, and who, if they can segregate two voices sufficiently, do not have to change position to switch their attention from one to the other.

There has been another attempt to segregate voices using only two microphones that were implanted in the ears of a dummy head. It employed a mathematical method known as adaptive noise canceling.[393] The computer processing was able to separate one target voice from a single interfering one almost completely and from three other ones quite well. However, the system had to be supplied with information about the direction of the desired speaker (by having him speak alone) before it could do this processing, and the position of the target voice or the microphones could not change thereafter. Furthermore, these results were obtained with a silent and nonreverberant environment.

The unnatural limitations on the recording situation that must be imposed when employing the existing engineering approaches to this problem suggest that if machine-based attempts at voice segregation were based on the way in which people do it, they might be more flexible. An attempt to model how the human auditory system may be using spatial location in speech separation was begun by Richard Lyon, who created a computational model in which the delay of information in one ear relative to the other was assessed separately for 84 different frequency channels.[394] It was based on a model for binaural location estimation that was proposed by Jeffress in 1948.[395] The channels in Lyon's computer model were meant to correspond with the output of 84 small regions of equal size on the basilar membrane of the human cochlea, covering the range 50–10 KHz. A comparison between the ears gave a separate apparent direction for each frequency channel. The estimation procedure was repeated every .5 msec. In many of the channels, the apparent direction did not correspond to any real sound source but was due to a mixture of sound energy from more than one source. However, the individual directional estimate for each small "sound fragment" (extending over a

narrow band of frequencies and short stretch of time) was dominated, more often than not, by the effects of a single source. Lyon had a restricted goal for this labeling, to use it to resynthesize the voice of one of two speakers, one to the left and one to the right of the listener. The way in which Lyon used these estimates after they were obtained reflected the limited goal that he had for it, but I see no reason why these estimates could not be combined with other sorts of information to decide how to connect Lyon's sound fragments over frequency and over time.

Interaction with Other Cues in Determining Grouping It seems that spatial origin, in itself, is such a good cue to whether or not spectral components come from the same sound, that we might need no other. Yet human auditory scene analysis does not put all its eggs into one computational basket. While spatial cues are good, they are not infallible. For example, they can become unusable in certain kinds of environments. One such case is a very reverberant environment. Another involves a situation in which the evidence for a particular sound is masked at one ear by a sound very close to that ear.

Even when the environment is more favorable, the mere fact that two auditory components have come from the same direction does not mean that they have originated from the same acoustic event. In vision, the surface of the nearer object occludes the sight of the further one, but this occlusion of evidence will not necessarily occur with sound. We have to remember that sound is transparent. Therefore two sounds that are coming at a listener from the same direction may mix their effects in the air. So when we hear two acoustic components coming from the very same place in space, we still cannot be sure that they came from the same acoustic event. Even a single frequency component arriving from one spatial direction could be the summation of corresponding frequency components arising from two different events. We do not know, in an ecological sense, how often the various cues are available or reliable in a random sampling of the environments in which humans find themselves. Very often, we suspect, the environment nullifies the usefulness of one cue or another.

For these reasons, the human auditory system does not give an overriding importance to the spatial cues for belongingness but weighs these cues against all the others. When the cues all agree, the outcome is a clear perceptual organization, but when they do not, we can have a number of outcomes.

When cues compete, the outcome may be determined by the extent to which the individual requirements of different cues are met. We can see how this works by looking at an experiment on the competi-

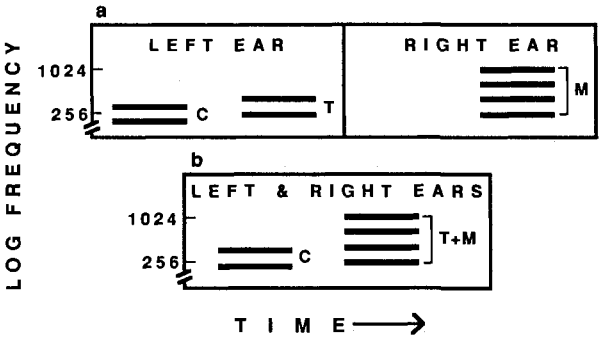


Figure 3.17
An example of a sequence of complex tones used as a captor (C), target (T), and masker (M) in an experiment by Steiger and Bregman (1982b). Presentations were either (a) dichotic or (b) binaural.

tion of cues done by Howard Steiger and myself.[396] The stimulus in this experiment consisted of a repeated cycle of two sounds, a captor tone, C, and a complex sound consisting of a mixture of two tones, T and M (see figure 3.17). The total length of one cycle was about 0.4 second. In the mixture (T + M), T was the target tone. The captor, C, was always very similar to T (same partials, each lowered by about 3 percent) and could pull T out to form a sequential stream, CT–CT–CT, and so on. M was the rest of the sound, which acted as a masker of T. Each of the sounds (C, T, and M) could consist of more than one partial.

As described so far, this stimulus offers the possibility of a competition between the sequential grouping of the captor and target (CT) and the spectral fusion of the target and the masker (T + M). However, there was also another type of competition: Sometimes the masker was in the ear opposite to the target and some partials in the masker could fuse with it to create a sound localized in the center of the listener's head. Therefore spatial cues sometimes entered into the competition. The listener's task was to adjust the intensity of the masker until a level was reached at which the target, T, was just audible.

The main purpose of the study was to find out which sort of fusion best opposed the sequential capturing effects of C. In other words, which masker, going on at the same time as T, best absorbed or masked it. The relations between the masker and the target were varied in three ways: harmonicity (relation to a common fundamental), number of corresponding harmonics, and same- or oppo-

site-ear presentation. We found that when the opposite ear heard a masker that *matched all the target's harmonics exactly*, the fusion tended to be strong and to prevent T from being segregated from the mixture. Under these conditions of an exact match, binaural fusion was much more effective in preventing segregation than the best case of ordinary spectral fusion was. However, when the match in harmonics was not precise, either because the harmonics in the masker were tuned to a different fundamental or were not precisely the same harmonics as in the target, monaural spectral fusion was stronger than fusion across the ears. In other words, binaural fusion appeared to be very narrowly tuned and to happen on a harmonic-by-harmonic basis. This finding supports the results that have been obtained in a different type of experiment on the binaural fusion of harmonics. These latter studies asked whether the harmonics would fuse across the ears to generate a single pitch or remain separate and be heard as two pitches, one at each ear. The results suggested that the across-ear fusion of pitches occurred only when every individual partial in one ear was close enough in frequency to its counterpart in the other ear to fuse with it.[397]

What does all this tell us about the functioning of the auditory system in real life? It says that the system expects a very precise match in frequency (within about 4 percent at frequencies below 1,000 Hz) between the harmonics at the two ears if it is going to have this match overrule alternative clues to the belongingness of acoustic components. The requirements are less stringent for the fusing of partials within an ear than across ears. This makes perfect sense in natural environments. Correspondences across the two ears will arise when the two ears are both hearing the same sound. If the sounds in the two ears are merely *similar*, this is not sufficient evidence that they are different views of the same event. Nature never sends the same sounds to different ears with their frequencies lowered in one ear. It is able to do so for loudness, but not for pitch. Therefore the ear should require a reasonably precise match in the frequencies of the components before it decides that it is hearing only one sound.

The reasoning is different when it applies to different partials received at the same ear. When these come from the same acoustic event, they correspond to one another in a different way than those that arrive at opposite ears. At opposite ears, it is possible for us to receive two separate registrations of the very same partial. In a single ear, each partial is represented only once; so when we hear two distinct partials, they are never just two snapshots of the same one. They may, of course, have arisen from the same acoustic event, but the conditions under which they should be accepted as such are not as

precisely defined as in the across-ear case. It may happen that two partials in the same ear are not exactly harmonically related, but what of this? Very often the partials in a natural sound may not be. So harmonicity is not a fixed requirement. The same pair of harmonics may have occurred together a bit earlier, but again this is not a prerequisite for accepting them as parts of the same sound. These ecological arguments may explain why within-ear fusion requires less exact relations between partials, but never ties partials too strongly together, and why across-ear fusion is so strong when its exact conditions are met.

The exactness of the requirements for across-ear fusion of speech sounds was demonstrated in experiments done by James Cutting, who sent a synthesized syllable "da" to both ears of a listener, but with slight inter-aural differences.[398] In one experiment, the two signals were identical except that one was delayed relative to the onset of the other. At delays of as little as 4 msec, the listeners began to segregate the signals at the two ears more often than they fused them. Remember that a syllable is a complex, time-varying signal; even though the "da" is synthesized on a constant fundamental frequency of 100 Hz (it sounds like a monotone), since the distribution of energy in the spectrum of the syllable is changing over time, as soon as one is displaced in time relative to the other, the spectra do not match exactly at the two ears. Cutting also did a version of the experiment in which the two "da" syllables were synthesized with different steady fundamental frequencies (pitches), but were otherwise identical. He found that with identical fundamentals there was 100 percent fusion of the two signals as one would expect, since the signals were then identical. However, with differences of as little as 2 Hz, which represents only one-third of a semitone at 100 Hz, the listeners almost always heard two sounds. In addition to being an illustration of the exactness of the requirement for spectral matches across the ears, this experiment may also be a demonstration of another point: although a spatial difference (having the sounds in two different ears) may not, in itself, segregate two signals, it may strongly assist other factors (such as a spectral mismatch) in doing so.

Contribution of Perceptual Grouping to Perceived Location So far, we have seen that conflicting cues can vote on the grouping of acoustic components and that the assessed spatial location gets a vote with the other cues. What is even more striking is that the other cues get a voice in deciding not only how many sounds are present but even where they are coming from in space. The auditory system seems to want to hear all the parts of one sound as coming from the same

location, and so when other cues favor the fusion of components, discrepant localizations for these components may be ignored. It is as if the auditory system wanted to tell a nice, consistent story about the sound.

We saw this earlier in some of the illusions generated by a rapid alternation of tones between the ears. Take, for example, Diana Deutsch's octave illusion, where there might be a 400-Hz pure tone at one ear and an 800-Hz tone at the opposite ear, with the positions of the two tones switching repeatedly at a fixed rate.[399] Many listeners will hear an illusion in which there is only a single sound, which alternates between the ears, following the position of the high sound, but whose apparent pitch alternates between high and low. Apparently this is the way that some listeners resolve a conflict of cues where (1) the harmonic relations vote that there is only one sound, (2) the appearance of the same sound alternating between the ears votes that every successive sound is part of the same stream or possibly the same event, and (3) the independent localizations for the frequency components vote that there are two separate sounds. If we eliminate the good harmonic relation between the two sounds by making the relation different than an octave, the illusion disappears, and the listeners' ears produce a more accurate account of what is going on.

Although we are always much more struck by the interaction of cues when they produce an illusion, we should be aware of the importance of this interaction in everyday life. As I pointed out earlier, comparisons of our left and right ear inputs are not always good cues for how to group spectral components. They are also not always very reliable in telling us where those components are located in space. For example, in chapter 7, I describe a synthetic pattern of sound in which identical frequency components presented to the two ears do not fuse to form a single sound in the center. Instead the co-occurrence is interpreted as an accidental correspondence of a frequency component from two different sounds. It is conceivable that the momentary binaural correspondence causes a centered location to be computed, but that the larger context in which it occurs corrects the interpretation so that we never hear a centered sound.

Less dramatic examples of the same type of correction must occur a hundred times a day. As our ears pass close to reflective or absorptive surfaces, and as different sounds start and stop, the classical binaural cues to localization must momentarily give incorrect answers and their short-term decisions must surely have to be compared to the previously computed description of the environment so that wild fluctuations of our perceptual descriptions can be prevented by some

conservative strategy for updating them. This would be a bad approach to take in a world in which each harmonic was on its own, not bound to acoustic events. In our world, however, harmonics and other products of our auditory system's analysis of acoustic events are treated as what they are: views of the same real-world events through different peepholes, where the glass in one or another peephole may momentarily be distorted or clouded.

We have seen, earlier in this chapter, that the apparent location of a sound in space can be corrected by the ventriloquism effect, the correlation of the behavior of the sound with that of a visual movement whose spatial location is different from the one that has been computed by the auditory system.[400] In this case, the typical perceptual result is the choice of an intermediate location for the sound. We think of this effect as an illusion, a view represented by the very name we give to it—ventriloquism. Yet it too is an example of the cleverness of our perceptual systems at resisting error by using the assumption of a coherent environment in which sounds should come from the places that are occupied by the events that made them. Once this assumption is built into our perceptual apparatus, it can correct errors in the estimates of the positions of events that it has obtained through sound by using estimates derived from vision.[401] Vision, of course, is not the ultimate criterion; so the final result takes both estimates into account. There is a problem, however, in the use of this procedure. When more than one event is occurring at the same time, which of the ones that have been located using vision should be treated as the same event as one that has been located using hearing? There are probably at least two criteria: one is that the two spatial estimates should not be too far apart, and the second is that the temporal patterns received by the two senses should show a weak synchrony. By "weak synchrony" I mean that not every event detected by one sense should necessarily be detectable by the other, but there should be a substantial number of correspondences. The cues might not be required to be absolutely synchronous but be allowed to be offset by some amount (perhaps up to 200 msec). When the perceptual systems decide that they are receiving information about the same event, they will merge information even when the auditory cues and visual ones are discrepant by as much as 30° (measured in terms of the visual angle).[402]

Interaction with Sequential Integration In an earlier section on sequential integration, I have reviewed a number of studies that show how sequential grouping by frequency can overcome spatial cues for the grouping of simultaneous sounds. I will now describe some experi-

ments in which the sequential cues were deliberately manipulated so as to segregate a part of a mixture from the rest and this affected the perceived location of that segregated component.

The first one, done by Howard Steiger and myself, involved cues for vertical localization.[403] Calling high pitches "high" is not just an arbitrary metaphor. Higher pitched tones actually seem to be coming from higher in space.[404] It has been suggested that this effect is related to normal spectral cues for vertical location where, because of the filtering effect of the pinna of the outer ear, the incoming waveform is modified differently depending on its vertical angle of arrival.[405] Whatever its cause we have a phenomenon in which the monaural spectrum is linked to localization.

The purpose of our experiment was to build a signal out of two parts, a white noise burst and a pure tone. If the tone was high in frequency and was fused with the white noise, it would color the noise burst so that it sounded high and was localized higher in space, but if it was low in frequency it made the noise burst sound lower with a corresponding effect on its localization. The trick of the experiment was to see whether capturing the tone into a sequential stream with other tones would cause it to no longer contribute its highness to the noise with which it was mixed. First we did an experiment to verify that high- and low-frequency pure tones would be localized as high or low in space even if they were presented over headphones. We deceived our subjects by placing two dummy loudspeakers on the wall in front of them, one above the other, and asked them to judge the vertical positions that each of a sequence of two pure tones came from. They were told that they would be hearing the tones over headphones at the same time but to try to ignore the headphone signal and to try to focus on the signal coming from the loudspeaker. It was believable to the subjects that they could be hearing sound from the speakers even with their ears covered by headphones because the headphone speakers were encased in a foam plastic pad that let external sounds through. In actual fact, there was no signal from the loudspeakers. The results showed that in the sequence, the higher tone seemed to come from a higher position than the lower one.

This cleared the way for a second experiment in which tones and noises were played together. In this experiment, the noises came out of one or other of the pair of loudspeakers and the tones were presented over the headphones, synchronous with the noises. The loudspeakers were hidden from the listeners by a curtain, but four vertical positions were marked on the curtain. The listeners were asked to judge the vertical location from which the noise had come and to ignore, as far as possible, the tones that, they were told, were there

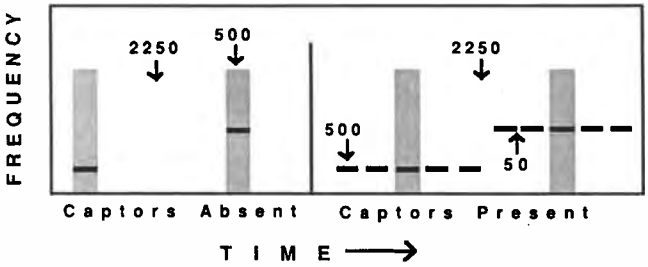


Figure 3.18
Stimuli used by Bregman and Steiger (1980). The vertical bars represent noise bursts and the dark horizontal bars are tones. The numbers represent the durations (in milliseconds) of the events or intervals.

only to distract them. A diagram of the stimuli is shown in figure 3.18. There were two conditions. In both of them, there was a sequence of two noise bursts synchronized with two pure tones of different frequencies. In one condition (captors present), each tone/noise burst was preceded and followed by two tones of the same frequency as the one that was present in the tone/noise mixture. These extra tones were intended to act as captors, capturing the tone into a separate stream from the noise. It was expected that under these conditions, the perceptual fusion of the tones and noises would be reduced and the tones would no longer determine the judged vertical positions of the noise bursts. The results supported this expectation: The direction of pitch change in the pair of tones affected the change in the judged position of the noise bursts only when the captors were absent and the tone and noise were allowed to fuse into a single stream. In a last condition, where there were no tones accompanying the noise bursts, the listener's judgments of their positions tended to be determined by the actual direction from which they had come.

This experiment showed how the fusion of simultaneous acoustic components can cause the spatial location information from them to be pooled into a single decision. Factors that cause the segregation of parts of this information, such as being captured into a simultaneous stream, can alter the perceived spatial properties of the auditory scene. Again we observe the effects of perceptual organization on the actual perceived location of sounds.

I have designed a simple demonstration to show that scene analysis can use sequential cues to influence localization in the left-right plane. It is created by playing the identical pure tone over channels A and B to stereo headphones. Then, using a volume control, channel B is repeatedly turned up and down between two limits—the intensity of

channel A and silence. When the intensity is raised and lowered slowly, the location of the tone seems to move from the A side of the head to the center and back again. This is what we would expect from our understanding of binaural cues for location. However, when B is raised and lowered rapidly, two tones are heard, one staying on the A side of the head and the other pulsing on the B side. The effect is enhanced by using briefer rise times for B, by keeping B at its full intensity for shorter periods of time relative to its time at zero, and by alternating the two intensities more rapidly.

The similarity of these conditions with those required for good illustrations of the continuity illusion (of a tone through noise) suggests that we are dealing with the same sort of effect. Although I will deal more thoroughly with the continuity illusion later in this chapter, let me anticipate that discussion by a brief word here. The present demonstration resembles the continuity illusion in that in both cases the sensory evidence, received when a signal is present, is partitioned into two parts, one treated as a continuation of the signal that precedes it and the other heard as an added sound. Both are examples of sequential integration overcoming simultaneous integration. There are two new features of the present example: First, the integration being overcome is binaural rather than spectral integration, and second, not only is the sensory evidence partitioned into two streams, but each is assigned a separate location. This example shows again that scene analysis can affect where we hear sounds to be and not just how many there are or what their auditory properties are.

In the examples examined so far, when there is a discrepancy between different cues that are used to calculate the location of sounds, the auditory system seems to adopt a few sensible strategies. In some cases, it allows one cue to win. In others it takes some sort of average between them. When the discrepancy is very large, the system may decide that two things have happened and may hear two sounds at different locations. In natural environments this third strategy is normally correct, but sometimes in the laboratory it can give rise to illusions in which things are heard at one location with properties derived from other locations. Some examples of this are the studies by Diana Deutsch and her colleagues on the octave illusion, the scale illusion, and the effects of a contralateral drone. We have already discussed these phenomena in an earlier section on the effects of spatial cues in sequential integration.

Conflicting cues can be used in a fourth possible way so as to generate the illusion of "duplex perception." Duplex perception can occur when different parts of a synthesized speech sound are presented to different ears over headphones. A sound presented in one ear plays

two roles at once. It can be heard as a sound in that ear and at the same time it alters the perceived identity of the sound that is heard in the other ear. Duplex perception has been thought to be such an unusual effect that it implies something special about the neural basis of speech perception. Although phenomena closely related to duplex perception can be observed with signals other than speech, I have decided not to discuss them now, but to do so in a section on exclusive allocation that follows the section on speech perception. This will allow me to address the apparent implications for the neurology of speech perception.

Interaction with Cues for Spectral Fusion The decision as to whether to merge information about spatial location from different spectral regions can depend on other sensory evidence that tells us that all these regions are registering the effects of a single sound. We have already discussed the demonstration made at IRCAM in which the even and odd harmonics of an oboe sound were sent to different loudspeakers. When the two sounds were encouraged to fuse, by giving them the same micromodulation, they were heard as coming from a single place, between the two speakers. When they were encouraged to segregate by being given different patterns of micromodulation, two separate localizations were made by the auditory system. This is another case in which scene analysis has controlled how classical cues to localization are allowed to operate.

A similar observation was made in an experiment by Donald Broadbent and Peter Ladefoged, in which two synthetic speech formants were presented to the same or different ears.[406] The formants could also be related to the same or to different fundamental frequencies. When the formants presented to different ears were related to the same fundamental, only one sound was heard and its location was more centered than in cases in which two formants with different fundamentals were presented. In the latter case two sounds with different positions were heard. The experimenters noted that the same factors that influenced the spectral fusion of sounds in one ear also controlled the fusion of sounds across ears. This supports the idea that all factors relevant to decisions about spectral parsing get to vote on it and when acoustic components are grouped into the same stream, for whatever reason, a single location is decided upon.

Conclusion: Classical versus Organizational Cues The most general conclusion that can be drawn from all these examples is that we must make some additions to the list of cues for spatial location that are typically given in textbooks on audition. Normally we would find

three cues. For left-right position there would be intensity and delay, that is the increased intensity of the sound in the ear that is pointed more directly at the sound source and the relative delay of the sound in the ear that is further away. For the vertical and front-back distinctions, certain properties of the shape of the spectrum that are affected by the shapes of our ears would be mentioned. We might call these "classical cues," but a better name for them is "ear cues." They are based on the shapes of our ears and how they are placed on our heads.

However, there is an entire other family of cues that we might call "world structure cues" because they have nothing to do with the placement of our ears. It is a property of the world, and not of our ears, that events create changes both in the light and the sound, and it is this property that is responsible for the ability to correct the location estimates that we derive from vision by using ones derived from audition and vice versa. Another important fact about the world is that sounds tend to last for a while and not to change very rapidly in their qualities. This fact make it very unlikely that if the sounds in a sequence have great acoustic similarities and are very close together in time, they are wholly unrelated. They probably came from a common source and should be put into the same perceptual stream. This justifies the system's adjusting the estimate for the location of each part of the stream by taking into account the estimated locations for the moments of sound that occurred immediately before or after it in the same stream. I do not claim that the human auditory system is actually doing anything like the logical calculations that I have just described. It simply puts together whatever it is built to put together. My statements about "justification" have merely described the utility of putting things together in this way. Presumably the utility is the same whether we are discussing people, earthworms, or robots. However, they may not all be built to exploit this utility. People are.

Other Factors Affecting Fusion of Simultaneous Components

So far, when we have discussed the segregation of spectral regions, we have typically spoken the language of Fourier analysis, in which the spectrum is filled with sinusoidal components. Alternatively, we have thought of the spectrum as filled with noise. We must, however, also consider sounds that are more like textures, the "granular" sounds that we considered in chapter 2 in our discussion of the acoustic factors that led to sequential integration. We mentioned sounds such as the crunching of the boots of a walker in snow, or the sound of something being dragged, or the riffling of the pages of a book. Analogous visual examples might include the hair of one's head,

where individual hairs are lost in an overall sweeping texture, or the appearance of a sheet of burlap.

In our earlier discussion, we found that since a certain amount of time was required to estimate the statistical description of the granularity, a difference in granularity would probably not be usable to create temporal boundaries, but might be usable for grouping separated sounds. Here we must do a similar analysis of the usefulness of differences in granularity for spectral integration. If we consider a spectrogram as a surface and if we looked at one that had a sufficiently high degree of resolution, these granular sounds might not look like individual spectral components (horizontal lines on the picture) or random noise (totally random dots) but might appear as different sorts of textures. If two spectral regions in this picture were filled with sufficiently different types of textures, the regions would visually segregate from one another.

There has been a considerable amount of research on the segregation of visual textures. One approach has been to describe textures as a distribution of grains on a surface where the grains have a limited number of shapes and the distribution has certain statistical characteristics.[407]

The question that we should address is whether different auditory textures in different frequency regions will tend to segregate from one another so that we hear two sounds with different qualities rather than only one. Unfortunately, no research has been done on this topic in audition. The critical step would be to devise some descriptive system for irregular sounds that could predict whether spectral regions filled with these sounds would segregate from one another. Then we would have to ask whether discontinuous parts of the spectrum (say in different frequency regions) would be integrated perceptually if they had the same description.

The problem gets more difficult if we ask whether differences in granularity could segregate two sounds when their spectra were overlapped. The granularity of the mixture will be the sum of the granularities of the two components and it might be very hard to recover the separate statistical properties with any degree of precision. Unfortunately we are left alone with our speculations on this matter. There is no evidence from the laboratory on the effects of granularity on spectral grouping.

We have now surveyed a large number of factors that affect how the auditory system carves up the immediately present spectrum and allocates parts of it to the perceptual representations of different events. I do not claim to have mentioned them all. For example, I have not mentioned the idea, put forward by Helmholtz in the pas-

sages that I quoted earlier, that different rhythmic patterns, present in different parts of the spectrum, indicate that there were different sounds embedded in it. Is this just an example of the use of synchronized onsets and offsets, factors that we have already examined, or does the periodic repetition of these synchronies make it easier to exploit them? There has been no research on this question.

All we can say is that there are a large number of factors that are useful in making this partitioning of the spectrum and are actually used by the human auditory system. We have seen that each cue has a partial validity in coming up with the right answer. For this reason I have called the auditory system's use of them "heuristic," a term in the theory of problem solving that means "tending to lead to good answers." What remains an open question is the exact method by which the auditory system combines the results of these heuristics to arrive at a decision that is usually right. There are researchers who are studying how computers could be made to recognize familiar acoustic patterns that might be present inside mixtures, and they would love to know the answer to this question. It might be very helpful if they, as specialists in system architecture, could suggest plausible ways that this might occur, so that students of perception could design experiments to determine which of these ways is used by the human perceiver.

Comparison between Fusion and Masking

This chapter has been describing how various factors promote the allocation of simultaneously present frequency components to the same stream. The evidence that I have used, in many cases, was that some components could be heard as perceptually separable sounds whereas others could not. This follows from the idea that when an acoustic component becomes absorbed into a stream that contains other simultaneous components, it gives up its perceptual identity in favor of contributing to the global features of the whole sound.

I now want to address the following question: What is the relation between perceptual fusion (absorption into a stream) and masking? I am prompted to ask this question for two reasons: the two phenomena are similar in their definitions (in both cases the listener loses the distinct qualities of one of the sounds), and the two phenomena seem to respond to the same variables in the same way. What then is the relation between perceptual fusion and masking?

The first thing to look at is the difference between their definitions. Helmholtz described our ability to hear a component in a mixture as analytic listening and referred to the tendency to hear the properties

of the mixture itself as synthetic perception. He thought of the latter as a less refined form of consciousness:

> We then become aware that two different kinds or grades must be distinguished in our becoming conscious of a sensation. The lower grade of this consciousness is that where the influence of the sensation in question makes itself felt only in the conceptions we form of external things and processes, and assists in determining them. This can take place without our needing or indeed being able to ascertain to what particular part of our sensations we owe this or that relation of our perceptions. In this case we will say that the impression of the sensation in question is *perceived synthetically*. The second and higher grade is when we immediately distinguish the sensation in question as an existing part of the sum of the sensations excited in us. We will say then that the sensation is perceived analytically. The two cases must be carefully distinguished from one another.[408]

I propose to examine how studies of perceptual fusion and masking relate to Helmholtz's ideas, but first I want to say a few words about the notion of analytic listening as a skill. It is surely the case that our ability to hear out components from mixtures is a learned skill whose execution may be deliberate and involve attention. Helmholtz himself became very good at hearing out several of the lower harmonics in a complex tone. Yet despite the fact that individuals may vary in this ability, I think there is a general rule that cuts across all cases. Everybody will find some components easier to hear out of mixtures than other components. This will depend on physical relations between the components of the sound and how our primitive processes of scene analysis can make use of these relations. So despite the presence of practiced skills, I think we will always be able to find evidence for the more primitive processes that depend very directly on the physical structure of the sound.

One might think that the studies both of masking and of perceptual fusion and segregation are looking at the listener's ability to do analytic perception of the target. Yet the measurement of this ability in the two types of experiments is quite different. In a typical masking experiment there are two sounds presented at the same time, a masker and a target. The listeners are supposed to listen for the target. The masker is made louder and louder until the target can no longer be detected. So far this sounds very much like a task involving analytic perception, and therefore resembles many of the tasks that I have described as showing the effects of perceptual fusion. However, in practice, the masking experiment is done a little differently. The

listeners are played a sound and must decide whether it consists of the masker alone or the masker accompanied by the target. Therefore they really do not have to detect the particular quality of the target sound in the mixture. All they have to do is to determine that the quality of the masker is different when a target is also present. The detection of this difference is not the same as hearing the target in the mixture, and does not imply that the target and mask are perceptually segregated. Masking studies, therefore, often measure the effects of the presence of the target on synthetic perception, because any global feature of the mixture can be used to make the discrimination.

When we say that the target has fused with the other sound, we are not implying that it will make no detectable contribution to the global qualities of the sound. On the contrary, the target might have fused with the masker but, in doing so, might have changed the quality of the masker. For these reasons, the typical masking experiment does not necessarily provide a measure of perceptual fusion. In a fusion experiment, on the other hand, the listeners are asked either whether they can or cannot hear the target in the mixture or, even better, to rate how clearly they can hear the target there. What we want to know is whether the target has retained its individual identity in the mixture. Ideally, they should also be told to ignore any qualities of the mixture that relate to *neither* the target nor the masker. (These would be emergent properties that arise from the combination of the two sounds.) There is an even better task to study the ability to perform analysis of the mixture. If a large set of recognizable and namable alternatives is used (such as words) the listener can be asked for the identity of the item. This task can clearly not be accomplished by basing one's judgment on some global property of the mixture. The study of the masking of speech, therefore, is not truly a study of masking but of perceptual fusion.

Despite the difference in the definition of fusion and masking, there are a number of results from experiments on masking that support the results on fusion. Variables that help to segregate one acoustic component from others seem to also prevent that component from being masked by the other ones. There are two possible explanations for this fact. The first possibility is that in many cases of masking, the listeners are actually trying to hear the target inside the mixture rather than just listening for a change in the global qualities of the mixture. By doing so, they are turning the task into one concerned with fusion.

The second alternative is that masking and fusion are different perceptual effects but that they both depend on the same underlying physiological phenomenon. For example, suppose a target is mixed with

a masker and the intensity of the target is gradually turned down. At first (let us call it the first stage) the target may be perfectly audible. Then, in a second stage, the target might not be audible as a separate sound but might still color the quality of the masker. In a final the masker might be heard as identical to how it appears when there is no target buried in it. The factors that favor segregation might act so as to make the physiological registration of the target stand out in some way so that it affects both the first and second stages of this sequence, requiring a greater strength of the masker to cause perception to move from one stage to the next.

An example that lends some validity to this description comes from the study of harmonicity. In a complex tone that contains many harmonics, one of the harmonics is gradually mistuned. Before it is mistuned the the complex tone has a single pitch and sounds like a single tone. With a large amount of mistuning, say over 6 percent, the mistuned harmonic is perceptually dissociated from the complex tone (i.e., no longer fused) and has its own pitch. However, with a small amount of mistuning, say from 1 to 5 percent, the change is detectable not as the appearance of a separate tone but as an alteration of the pitch of the complex tone. In the latter case, if we asked whether the complex tone were masking the mistuned partial the answer would be no, because the complex tone that contained it would have a different pitch than one from which it had been removed.[409] This demonstration of the effects of violations of harmonicity on both masking and perceptual isolation may arise from its action on a common mechanism.

For whatever reason, results from the two types of experiments tend to agree; so let me proceed to review some results from masking experiments that show the influence of factors that I have already described as being used by the auditory system to partition the spectrum into separate streams.

The first experiment that I want to describe showed that the segregation that comes from having different patterns of micromodulation can also affect how strongly a loud sound masks a weaker one. In 1978, a Dutch researcher, R. A. Rasch, of the Institute for Perception (TNO) in Soesterberg, reported a study of this effect.[410] He played subjects a sequence of two chords. Each chord consisted of two harmonically rich tones. Each tone had 20 harmonics with successively higher harmonics attenuated more and more so that a harmonic an octave higher than another one would be 6 dB less intense. The lower tone was always the same in both chords but the higher tone was different, either higher or lower in the second chord. Each tone was 200 msec long. The listeners had to say whether the high tone moved

up or down across the pair of chords. Typically the low tone was much louder than the high tone and acted to mask it.

Before describing the results, we should observe that despite the fact that the experiment is often described as a study of masking, according to our earlier definitions it is really about analytic listening or its opposite, perceptual fusion. This follows from the fact that the listeners had to make a decision about a quality of the target itself (its change in pitch) and not just a quality of the mixture.

In some conditions of the experiment, Rasch applied a simple frequency modulation (vibrato) of 4 percent to the high tone, the modulation being sinusoidal in form and repeating at 5 cycles per second. The low masking tone, however, was always steady in frequency. He found that the high tones were harder to mask when they underwent frequency modulation than when they were steady. If we were to set the low tone to a level loud enough to mask a steady high tone and then started to modulated the frequency of the high one, the low tone would no longer mask it and we would have to crank up the intensity of the low tone by about 17 dB (about a seven-fold increase in amplitude) to restore the masking effect. It is unfortunate that Rasch did not compare the effectiveness of the masking tone when it was modulated coherently with the target tone to the case in which it was modulated in an uncorrelated way. As the results stand now, I can think of two possible explanations for them: The first possibility is that the presence of two different forms of pitch modulation (sinusoidal for the target and nothing at all for the masker) improves the segregation of the two tones and thereby prevents the weaker from losing its identity in the mixture. This is what I would like to believe. However, there is a second logical possibility. Frequency modulation of a tone may improve the definition of its pitch, perhaps because the auditory system tends to habituate to steady tones, and this might make it harder to mask, even by a tone that was changing in parallel with it. Until experiments are done with frequency modulation of both masker and target, we will not be able to tell whether FM-based segregation is the cause of the release from masking that Rasch observed.

A second experiment was also done by Rasch as part of the same series. This one showed that asynchrony of onset prevented masking. In some conditions the target tone started a bit before the masker but the two ended at the same time. This made the target easier to hear. Each 10-msec increase in the time interval by which the target's onset preceded that of the masker made the target as much easier to hear as increasing its intensity by about 10 dB. The listeners were not aware that the onsets of the two notes were asynchronous; they knew only that they could hear the target better in these conditions.

Rasch was concerned about whether the listener was just basing this increased ability to hear the high tone on the few milliseconds of it that preceded the onset of the masker. To find out, he included a condition in which the target came on before the masker, but when the masker came on, instead of allowing the target to continue along with it, he allowed it to accompany the masker for only about 30 msec and then shut it off. If the listeners' detection of the high tone depended only on the part that preceded the onset of the masker, their ability to perceive the target should not have been affected by shutting it off in this way. And indeed it was not. This implied that they were not really segregating the spectrum of the target tone from the complex target-plus-masker spectrum but simply getting a glimpse of the target when it "stuck out" of the mixture. The listeners, however, thought that the high tones had lasted throughout the duration of the masker.

Before concluding that onsets or offsets that occur inside a mixture have no segregating effect, we should remember that in Rasch's study the masker tone was never *turned on* inside the mixture. That is, he never tested asynchronies where the target came on second. Michael Kubovy has shown that changing the intensity of a component tone inside a mixture of tones makes that tone stand out of the mixture.[411] Also Scheffers, in research that will be discussed in chapter 6, found that in mixtures of vowels, turning on the target vowel a few tenths of a second after the masking vowel had begun made the target much easier to hear than if the two started synchronously.[412]

Incidentally, a study of asynchronies by Dannenbring and myself gave a result that tends to support Rasch's conclusion that only the part of the target that sticks out of the mixture affects its perception.[413] We found that onset and offset asynchronies of a target harmonic relative to two accompanying harmonics made this harmonic easier to capture into a sequential stream, but this advantage seemed to be restricted to those cases in which the target tone stuck out of the mixture. Probably this was due to the fact that when the target went on or off when the rest of the mixture was on, the onset or offset was masked by the nearby onsets and offsets of the other tones as well as by their mere presence. Judging from Kubovy's results, if the onset of the target had occurred further inside the mixture, it probably would have escaped these masking effects and the target would have become audible.

There is one final point on the topic of onset synchrony that is of interest to musicians. Rasch pointed out that in polyphonic music the asynchronies of onset that occur because of the normal variability in

human performance would be sufficient to isolate notes from one another despite the fact that they were nominally synchronous in onset. He verified this claim in a later experiment that directly studied how exactly musicians synchronized their playing in an ensemble.[414] He found that nominally synchronous onsets typically had a standard deviation of differences in onset time ranging from 30 to 50 msec.

One other observation shows that it is easier for a masking sound to mask a target when both come on at the same instant (by "easier" I mean that the masker does not have to be so close in frequency to the target to mask it). Masking is harder when the masker is present continuously and the target comes on periodically than when both the masker and the target are gated on at the same time.[415] This is consistent with our observation that onset synchrony tends to fuse two sounds into a larger spectral pattern.

Comodulation Release from Masking

There is a phenomenon that shows an effect on masking that is similar to the effects of common amplitude modulation that I described in an earlier section. In that section, we saw that two simultaneously presented tones tended to segregate from one another perceptually if they were amplitude modulated at different frequencies or if their amplitude modulation was out of phase. The related masking effect is called comodulation masking release. It was discovered by Joseph Hall, Mark Haggard, and Mariano Fernandes of the MRC Institute of Hearing Research in Nottingham, England.[416]

Let me describe it. The detection threshold was measured for a 400-msec, 1,000-Hz pure tone when it was masked by different types of noise. In one set of conditions, the masker was a band of random noise centered on the frequency of the target tone. In different conditions, the bandwidth of this noise was increased, while holding the spectrum level constant. What this means is that the bandwidth was increased by adding new energy in the sections of the spectrum on both sides of narrower band without altering the energy in that narrower band. As energy was added and the noise band became wider, the target became harder to detect. However, after a certain width was reached, the interference got no worse when the band was made even wider. The existence of an upper limit derives from the fact that only the energy in a spectral band adjacent in frequency to a given target is effective in masking it. This region is called the critical band. If you add energy outside it, it has no further effect on masking. So far, then, the results were entirely predictable.

There were, however, a set of conditions in which the masker was not a simple noise band but a noise that, after being generated, had

been amplitude modulated in a random pattern.[417] Because of this way of generating the signal, the intensity fluctuation in the noise occurred in much the same pattern in all the frequency bands contained within the noise. As you increase the bandwidth of this type of modulated noise, you are adding energy that has the same pattern of amplitude fluctuation that is present in the narrower band. This is quite different than what is true in ordinary random noise, in which the intensity fluctuation is independent in each frequency band. There was a corresponding difference in what happened to masking when this type of noise was used as a masker and was varied in bandwidth. As with the unmodulated noise, the amount of masking rose as the noise band was made wider, up to the limit of the critical band. But as the bandwidth of the masker was made even wider, instead of remaining the same, the masking began to *fall* in effectiveness. In some paradoxical way, adding more energy to the masker was reducing its ability to mask.

The scene-analysis explanation of this result reduces to the statement that "it is easier to protect oneself from an enemy whose characteristics are known." The experimenters argued that because of the fact that the amplitude fluctuation was similar in different spectral regions, the auditory system was able to integrate the fluctuation knowledge from across the spectrum and use it to better cancel its effects out of the signal. It is possible to look at it a little differently, in a way that relates it to spectral grouping. The experimenters' explanation makes an assumption that we also have had to make in order to explain the segregation and fusion of amplitude-modulated tones, namely that the amplitude variation occurring in different parts could be separately detected and compared. Only in this way would amplitude fluctuation be useful in deciding which parts of the spectrum should be put together into the same stream. Following that same line of explanation in the present example, we would say that because the same amplitude fluctuation was detected in a number of frequency bands, those bands were linked together into a common stream. As more frequency bands were added, each containing a similar pattern of amplitude variation, the tendency to reject a band that did not have this pattern grew stronger and the target became more audible. This form of the argument shows explicitly the relation between the present effect and the general process of scene analysis.

Although this was discovered in the laboratory, the people who discovered it were aware of its importance in daily listening environments. I have quoted their argument earlier. I shall do so again:

> Many real-life auditory stimuli have intensity peaks and valleys as a function of time in which intensity trajectories [changes] are highly correlated across frequency. This is true of speech, of interfering noises such as "cafeteria" noise, and of many other kinds of environmental stimuli. We suggest that for such stimuli the auditory system uses across-frequency analysis of temporal modulation patterns to help register and differentiate between acoustic sources.[418]

This experiment is also relevant to the "peek" theory of the effects of amplitude modulation that I discussed earlier. According to that theory, the reason that it is easier to hear a target sound when it is present at the same time as another one that is being amplitude modulated in a different pattern may not be directly related to any strategy for parsing the spectrum. The auditory system may simply be getting a better peek at the target component at those instants of time at which the others have been reduced in intensity by the modulation. There is some experimental evidence to support this theory.[419] Still, the peek theory could not explain the present results. In particular, it cannot explain why the peeks should get more effective when the bandwidth of the masker increases.

An additional finding about the comodulation masking release phenomenon is that it does not work with frequency fluctuations.[420] If the target tone is masked by a signal that is varying in frequency in a random pattern, you cannot reduce the masking by adding a second signal that is varying in frequency in the same pattern and that lies outside the critical band of the target. This result may shed light on a question that was raised earlier in this chapter: Could the apparent integrative effects of parallel frequency modulation of subsets of partials be attributed entirely to other causes, leaving nothing over to be explained by "common fate in FM" principle? The comodulation data suggest that the answer could be yes.

There is another paradoxical masking effect in which adding more energy to the masker reduces the amount of masking. This one, too, is susceptible to a scene-analysis explanation. The effect occurs in forward masking in which a loud noise burst masker precedes a fainter pure-tone target. The masker is a band of noise centered at the same frequency as the target. If the bandwidth of this masker is increased, there is less forward masking. It has been argued that this occurs because a wide-band noise does not sound as much like the target as a narrow band of noise does. Therefore, a wider bandwidth of the masker will reduce forward masking because you can tell more easily where the noise ends and the signal begins.[421] Notice that this ex-

planation has the flavor of a scene-analysis explanation. It is not really in the tradition that attributes masking to the swamping out of the target information by the masker energy within a critical band. In that tradition, nothing that happened outside the critical band could have an influence. The explanation that says that the additional energy makes the masker more distinctive is really talking about the fusion of the various spectral regions in the wide-band noise to create a larger perceptual entity that has global properties that are different from the properties of its parts. Furthermore, it no longer talks about the swamping of the target by the local energy of the masker, but about the degree to which the auditory system integrates or segregates the target and the masker. I would suspect that when the process of spectral integration fuses the spectral components of the noise into a single sound, it has two beneficial effects. As well as computing a global quality for the noise, it integrates the simultaneous "energy-going-off" information across the spectrum that signals the turning off of the masker; it may therefore be able to get a more precise fix on when the target turned off. This may aid it in calculating that there is one frequency band (the one occupied by the target) at which energy did not turn off at the same time and that therefore there is something different going on at that spectral location.

There is also evidence that when a target and a masker are perceived to be at two different spatial locations, masking is weaker. For example, when a listener is presented with a voice that is being masked by noise, the intelligibility of the speech signal depends on whether the speech and the noise are coming from the same spatial location. With increasing separations the speech becomes more intelligible.[422] Actually many of the studies have manipulated the interaural time delay rather than the actual position in space of the signals.[423] Masking drops off with increases in the spatial separation between target and masker.

This spatial release from masking can be seen in a phenomenon known as the binaural masking level difference (BMLD).[424] To illustrate it, we begin by presenting, over headphones, a mixture of a tone and a noise binaurally (i.e., with identical copies to the two ears). Then we increase the intensity of the noise until the tone is no longer audible. That is stage 1. In stage 2, we simply shift the phase of the tone in one ear, so that it is now out of phase by 180° across the ears while leaving the noise identical in the two ears. The tone will now become audible again. We will need to boost the intensity of the noise again in order to mask the tone. The difference between the intensities of noise needed in stages 1 and 2 to mask the tone is known as the binaural masking level difference.

Another example of the same phenomenon is also accomplished in two stages. In the first, we present a mixture of the tone and noise to a single ear and increase the intensity of the noise until it just masks the tone. In the second stage, we keep the first-ear signal as it was, but we now present a signal to the opposite ear. This signal contains only the noise, but as soon as we present it, the tone at the first ear becomes audible.

It is tempting to conclude that the results are due to the fact that phase relations between the two ears are a cue to spatial location. In all cases of the BMLD effect, one signal (target or masker) is in phase at the two ears and the other is out of phase. Therefore the two signals should appear to have different spatial locations. If the auditory system segregated them on this basis, this might make them less likely to fuse and would make the target more audible as a distinct sound. Indeed, in many examples of the effect, such as the second example that I gave, the sounds are heard as having different locations.

However, in cases where there is a reversal of phase between the signals in the two ears, as in the first example that I gave, the phase difference is greater than what would occur in natural listening situations and the localization of the phase-reversed sound is diffuse. Therefore it appears that a clear localization of the target and mask in different places is not required for obtaining the BMLD. The only thing that seems necessary is that the between-ear phase comparison should give a different result for the target and the masker. This requirement does not destroy an explanation based on scene analysis, only one based on a scene analysis that occurs after separate spatial locations are assigned to target and masker. We have seen that the auditory system can make use of any one of a number of acoustic relations to decide whether or not two parts of the spectrum should go together. Perhaps one of them is interaural phase relations. If two parts of the spectrum come from the same source, then they should be coming from the same spatial location and be subject to the same echoes; therefore, the delay between time of arrival in the two ears should be the same for both parts. Working this reasoning backward, if the auditory system receives two spectral regions with the same interaural phase difference, it should fuse them; however, when the two regions show different interaural phase differences, the regions should be segregated.

Perhaps this heuristic can work even in cases where the spatial location of one or the other sound is ambiguous. Even though location cannot be fully decided on when the phase is reversed in the two ears, it can still be decided that one part of the spectrum has a different interaural phase than another part. Therefore phase relations may be

able to provide segregation of identity even when they cannot give definite locations. Earlier in this chapter, I said that separate assessments of spatial location had the power to segregate and group parts of the spectrum. However, what I took to be decisions about spatial location might have actually been something that is a simpler precursor of spatial localization, namely separate assessments of interaural phase relations.

I would not like to argue that it is the phase relations alone, and never the spatial estimates themselves, that are used to parse the spectrum. There are other cues to spatial location besides interaural phase relations. For example, there is the rule that a sound should not change its location too rapidly. There is evidence that these other cues can also affect the parsing of the spectrum.

To sum up the evidence from masking experiments: We see that masking and fusion seem to be affected by the same sorts of acoustic relations between a target component of a larger sound and the remainder of that sound. Generally speaking, these are relations that, in a natural listening environment, are useful for deciding whether different spectral components have arisen from the same acoustic event. Despite the differences in the measurement operations that define fusion and masking, the two may be affected by common underlying mechanisms.

Meaning of These Findings for Biology It is interesting to step back and look at the scene analysis heuristics from a biological perspective. We have looked at a number of relations between simultaneous components of a sound that are unlikely to have occurred by chance and therefore offer the auditory system an opportunity to exploit them for scene analysis. We have found that the auditory system seems to have evolved ways of taking advantage of them so as to make the right combinations of components hang together for purposes of more detailed analysis. We might ask at this point whether these findings imply anything about what physiologists should look for in the neural architecture of the auditory system. I am afraid that I am not very hopeful about drawing inferences from function to architecture. True, the auditory system does exploit these regularities. However, we should realize that a biological system might not go about this in the way that a careful computer designer might. We may not find a specific neural computation, for example, of the frequency change within each frequency band and then a second computation that sorts the bands into groups that show similar changes. As an example of the indirect methods of the auditory system, we have already discussed ways in which a peripheral auditory mechan-

ism might turn a physical frequency fluctuation into a neural registration of an amplitude fluctuation. Or we might find, for example, that the reason that partials with synchronous onsets tend to fuse is that their onsets tend to mask one another and therefore cannot be heard separately.

Scene analysis, as I am discussing it, is a function, not a mechanism. It "takes advantage of" regularities in the input; it may never directly "mention" them. We may never find a single mechanism that does nothing but scene analysis. When nature evolves something, it often does so by using materials that were already in place for some other reason. The fins of the whale, for example, are adaptations of earlier legs, which, in turn, are adaptations of yet earlier fins, and so on. As I pointed out in chapter 1, neurological breakdowns may be the mechanism underlying functional accomplishments. We should not expect a direct mapping from function to mechanism. It may exist in some cases, and in those cases, science is fortunate. We should therefore not be dismayed if many of the accomplishments described in this chapter are found to be incidental byproducts of some well-known mechanisms, for this finding would not imply that they were not specifically evolved for scene analysis. Evolutionary selection favors the mechanism with the useful by-product over one that lacks it. Thinking in terms of scene analysis may not lead us to find new neural mechanisms. But it may help us to understand the purposes of the ones we know about.

Perceptual Results of Simultaneous Integration and Segregation

The last section was concerned with the factors used by the auditory system to decide which simultaneously present sensory components should be assigned to the same perceptual stream. The next section will discuss what we know about the use that is made of these grouped elements. Before even starting to look at the results of experiments, I would suspect that every analysis that can be made on a group of components heard in isolation can also be made on a subset that has been segregated out of a larger set: When we listen a particular sound in everyday life it is rarely the only one present and yet we seem to be able to assess its loudness, pitch, timbre, distance, direction, and so on. Many of these features are the results of analyses that must take into account a set of grouped acoustic properties. They pertain to the sound as a whole and not to its individual components. For this reason they can be called global properties. For example, the global pitch of a sound depends on the analysis of only those partials

that have been grouped together. Presumably the global timbre does likewise.

In chapter 1, there was some discussion of emergent properties, properties that do not apply to the parts of a thing, but to the thing as a whole. A visual example might be the property of being a closed figure. Suppose we have a figure drawn using straight lines. None of the individual lines in it is a closed figure and descriptions of the individual lines will not be phrased in terms of this property. The property emerges when the interconnections of the lines are taken into account. It is a property of figures, not of lines.

When I speak of global properties, I am thinking of a closely related idea. For example, I have suggested that a global analysis of a fused set of partials might find a pitch appropriate to the set of partials as a whole, whereas if the partials were perceptually decomposed from the mixture, the partial pitches could be heard. Pitch is not a property that applies only to complex tones. It can be heard for partials as well. Therefore the word "emergent" is not quite right to describe the pitch derived from the whole set of partials. We need a word that means that a certain property, such as pitch, has been computed from all the components that are present and not from a subset. I have chosen the adjective "global" for this purpose.

The common idea behind the two concepts is that they describe features that have been computed on some grouped set of components. If the feature is emergent, it means that the lower-level components could not have this sort of property; an example is a closed figure. If the property is global, it means that the property was based on a grouped set of components, but there is no implication as to whether the lower-order components could have this type of property, an example being a global pitch. Since "global" is the more inclusive word, I will try to use it in the present discussion.

I sometimes use the word "partial" in expressions such as partial pitches, partial timbres, and the like. By this I mean to refer to the fact that less than the total number of acoustic components present are contributing to the perception of that property. By the term partial pitch, I am also usually referring to the pitch of a particular partial, since that is often what we can hear in an inharmonic tone.

Although the previous section focused on the factors that promoted the segregation of simultaneous frequency components, it was not possible to discuss this topic without mentioning some of its effects. After all, the only way to study the causes of segregation is to know that a segregation has taken place, and this requires us to at least have a good guess about the perceptual effects of such segregation. Therefore, rather than introducing something entirely new, the

following section will simply try to list in one place the various perceptual effects that have been observed when a complex sound mixture is segregated into co-occurring streams by the auditory system.

Examples of Within-Stream Computation of Properties

Within-Stream Temporal Properties; Streaming Rules The most obvious effect that can be observed is that the mixture can be heard as two or more event sequences, each having its own independent temporal properties. A familiar example of this occurs when two voices are heard at the same time. The fact that we can often focus on just one of these and understand it implies that the sequential properties of each are kept distinct from the other. The interpretation of speech depends on the order of more basic speech events (perhaps phonemes) that occur within it. Therefore, if we can interpret the message of each voice separately, we must be able to form a sequence for that voice that excludes the other sounds.

An even more obvious case of the necessity for being able to register the within-stream order of events is in the recognition of melodies in music. Here we can fall back on data from an experiment that shows that it is possible to recognize a melody whose notes are being played in synchrony with the notes of two other melodies. In this experiment, the melodies were arbitrary arrangements of four pure tones, in the vicinity of 500 Hz, with the interfering melodies being more or less symmetrically placed, one higher and one lower than the target melody in frequency. The melodies were presented at a rate of 2.5 tones per second. The listeners had to distinguish between the test melody and an alternative one under conditions in which either one could be present as the mid-frequency melody in the three-melody mixture. The alternative melody contained the same tones as the test melody except that the temporal order of the middle two tones was reversed. In different conditions the masking melodies were at different frequency separations from the target. It was found that the melodies became easier to recognize as the frequency separation between the target and interfering melodies was increased. This recognition, moreover, clearly must have depended on the correct detection of the order of tones in the target.[425] The effect of frequency separation in promoting the segregation of different streams of tones resembles the effects of this variable that we saw earlier in our examination of the streaming phenomenon in rapid sequences of tones. The only difference was that in the stimulus pattern for the streaming phenomenon, the tones were never physically overlapped in time

even though the listener sometimes interpreted the streams as existing at the same time.

Pitch Another property that depends in some way on the integration of simultaneous components is pitch. We have seen how the micromodulation in synchrony of a set of frequency components can isolate those and cause them to be the ones that contribute to the pitch analysis. An example of this that we have not yet discussed was provided by Stephen McAdams, who carried out an experiment in which he synthesized a mixture of three sung vowels, each on a different fundamental frequency.[426] Without any micromodulation, the listeners sometimes heard four to six pitches. But when the sounds were modulated, they tended to hear the three pitches, one for each fundamental. This suggests that the micromodulation might have encouraged the auditory system of the listener to put all spectral regions of the signal into a global pitch analysis. McAdams also reported that several other authors have noted that the "missing fundamental" pitches derived from certain stimuli are better heard when the complex is modulated coherently than when it is not modulated.[427] These all represent cases in which spectra are integrated by scene analysis so that a global pitch can be computed.

We have to be careful about generalizing from the results using micromodulation. So far I am aware of no demonstration that any other factor that promotes the grouping of components can affect the pitch computation. A plausible one to try would be the spatial separation of the components. Again, care must be taken in interpreting results. For example, it has been shown that two successive upper partials can be presented, one to each ear, and they will be perceptually grouped across the ears for the purposes of calculating a missing fundamental. This perception of global pitch is clear enough to allow listeners to recognize the musical interval formed by the fundamentals in a sequence of two such presentations.[428] This, however, does not show that spatial separation has no effects whatsoever on the pitch calculation. As we have seen, pitch is one of those properties that exist either for individual components of for subsets of components. In some cases, both the global and the partial pitches can be heard at the very same time; under some conditions the global pitch is strong and the partial ones weak and under others the opposite is found. The dominance of the global pitch over the partial pitches is not an all-or-nothing affair. Therefore in the experiment that I just described, we do not know whether the across-ear integration was weaker than it would have been if the pair of tones had been presented to a single ear. Actually the perception of the missing fundamental

was fairly weak, and some of the listeners, despite previous musical training, had to be given up to an hour of training before they could hear it. In my own listening, I am never quite sure that I can hear the fundamental with such stimuli. To really know whether spatial separation reduces the integration of components for pitch computation, it would be necessary to do a study in which the strengths of the pitch percepts were compared in the across-ear and the within-ear cases.

Timbre What of timbre? We would expect two separate timbres to be able to be heard when two subsets of partials were segregated out of a mixture. We have seen one example of this in the sound that was created for musical use at IRCAM. When the odd and even harmonics of a synthesized oboe sound were segregated by location and by different patterns of micromodulation, the quality of the two separate sounds was not the same as the global oboe timbre that was heard when they were not segregated. Another example of scene analysis affecting the perception of timbre was mentioned when we discussed spatial location as a factor influencing timbre. There we saw that when one partial was sent to one ear and the rest of the partials to the other one, the listeners were not able to form the two into a single "spectral profile."[429]

Richard M. Stern has reported a related experiment on timbre perception. Synthetic trumpet and clarinet tones were presented in pairs, either at the same pitch or at different pitches.[430] When the two were at the same pitch the listener heard a single note that sounded neither like a trumpet nor a clarinet. But when the two pitches were different, a musically trained listener could tell which instrument was playing which note. Evidently the segregation of two sets of partials that were harmonically related to different fundamentals allowed an independent computation of timbre to take place on each set.

Vowels A perceptual property similar to timbre is the quality that distinguishes one spoken vowel from another. Both the distinction of one vowel from another and one timbre from another depend on the pattern of the amplitudes of different spectral components. If vowel quality is seen as an example of timbre, then research on the factors that makes two simultaneous spoken vowels distinguishable from one another is really research on the segregability of timbres. I plan to describe this research later when I deal more extensively with speech perception. For the moment it suffices to say that differences in fundamental frequency between the partials of two simultaneous vowels makes it much easier to identify the vowels.

In a way, research that uses vowel sounds as examples of timbres is easier to do than research with arbitrary timbres. It has the advantage that the listeners already have names for the sounds they are hearing and, for this reason, experiments that ask for identification of the components sounds are easier to set up. It has one disadvantage though. Since speech perception is highly overlearned, and may also make use of an innate mechanism that is specialized for speech, we cannot be sure whether the segregation that we observe is due exclusively to the basic scene-analysis capability of the auditory system. The only way to decide is by seeing whether the same set of factors that is known to affect the fusion or decomposition of other nonspeech mixtures also affects mixtures of speech sounds. We will take up this issue in more detail in chapter 6.

Consonance and Dissonance Musical consonance and dissonance can be classified under the general rubric of "timbre" as well. It arises, acoustically, from the beating of certain partials in the mixture of two or more tones. Later, in chapter 5, we will see how the experience of dissonance that arises when two sounds are played together can disappear when the two sounds are segregated into separate streams.

The apparent ability of the auditory system to not hear the dissonance presents us with a theoretical dilemma. Does the auditory system not register the beating that occurs when two acoustically dissonant tones occur together? After all, beating is a phenomenon that is physically based. Surely the auditory system cannot overcome the physics of sound. What I am claiming is that although the beating of the partials of the two tones may be registered at some peripheral level of the auditory system, it is not assigned as part of the mental description of either sound because the auditory system can somehow tell that it is not an intrinsic property of either sound but, rather, an accident of their combination. It has been claimed by workers in visual information processing that the conscious experience of disembodied features is impossible. Features have to be assigned to a percept in order to be perceived.[431] Perhaps acoustic interactions between simultaneous events are picked up but not assigned to any stream and therefore remain outside the level of awareness.

In many preceding discussions I have described the allocation of a piece of spectral evidence to one stream or another as if this were an all-or-nothing effect. While it was convenient to talk that way for purposes of exposition the actual facts are not as clear cut. When we listen to a mixture of sounds, we may be able to perceptually segregate one of the components from the mixture, but our experience of

this component is not exactly as it would have been if the segregated component were not accompanied by the other ones. The experience is somehow less clear and features of the remainder of the sound often "leak" through. This failure of all-or-nothing allocation means that some features are playing a double role: They are contributing to the description of the sound that we are focusing our attention on, but they would also contribute to the description of the background sound if we were to focus our attention on that one. It appears, then, that the principle of exclusive allocation does not operate as strictly as I implied in chapter 1. The failure of this principle leads to a phenomenon called duplex perception, in which a piece of acoustic evidence is used to construct two different percepts. Because this was first noticed in the context of speech research, it was taken as telling us something very special about speech perception. It is therefore convenient to discuss this whole issue in chapter 7, which considers the violations of exclusive allocation in speech perception.

Not all segregations are equally strong. We are not always equally capable of segregating a sound from a mixture. Some of this depends on our learned skills, but very often the signal itself imposes limitations on what we can do with it. These limitations come in two forms. One is related to certain limits on the resolving power of our peripheral apparatus. If two tones have frequencies that are too close together, rather than perceptually segregating the two tones, you will simply hear the pattern formed by their sum. A second limitation relates to whether or not the cues for segregation are unambiguous. Clarity will occur when a number of cues all suggest that there are two distinct sounds whose descriptions should be created independently. When some cues point in the other direction (as would happen, for instance, when two unrelated sounds happened to start and stop at precisely the same time) the segregation at the preattentive level would be less strongly created and this would make it harder for later processes, such as selective attention, to treat the two sounds as distinct.

What Is the Default Condition: Fusion or Decomposition? In this chapter, we have looked at a number of factors that lead the listener to either fuse or segregate a set of simultaneous components. Which is the default condition, fusion or decomposition? One could imagine two states of affairs. In one, the auditory system prefers to fuse the total spectrum at any moment and is prevented from doing so only if there is specific information that parts of it should be segregated. The alternative is that the auditory system tries to treat each frequency

region separately unless it has evidence that it should group some of them into a more global sound. Is the total spectrum assumed to be innocent of internal structure until proven guilty or is it the reverse?

An observation made by Helmholtz is relevant to this point. He reported what happened when, using tones produced by bottles, he used a two-tone complex and started one tone (the higher) first. At first he could hear the upper partial, with its characteristic timbre clearly. Then it began to fade and become weaker. The lower one started to seem stronger and dominated the timbre, and the upper one started to sound weaker. A new timbre began to emerge from the combination. According to his description, he could hear out an upper partial of a complex tone whose two components were spaced by an octave; however,

> . . . I could not continue to hear them separately for long, for the upper tone gradually fused with the lower. This fusion takes place even when the upper tone is somewhat stronger than the lower. The alteration of the quality of the tone which takes place during the fusion is characteristic. On producing the upper tone and then letting the lower sound with it, I found that I at first continued to hear the upper tone with its full force and the under tone sounding below it in its natural [context-free] quality of *oo* as in t*oo*. But by degrees as my recollection of the sound of the isolated upper tone died away, it seemed to become more and more indistinct and weak while the lower tone appeared to become stronger, and sounded like *oa* as in t*oa*d. This weakening of the upper and strengthening of the lower tone was also observed by Ohm on the violin. . . . With the tones produced by bottles, in addition to the reinforcement of the lower tone, the alteration in its quality is very evident and is characteristic of the nature of the process.[432]

This observation seems to imply that the default is fusion since when the segregating effects of the asychronous beginning of the two sounds receded into the past, the tones tended to fuse. However, this interpretation fails to take into account that there was not only a cue for segregation in this situation but one for fusion as well. The cue for fusion was the harmonic relation between the two sounds, the upper one being an octave above the lower one. This would make the upper one the second harmonic of the lower. Therefore the asynchrony cue was not acting alone but was opposing a cue for fusion. Furthermore although the segregation cue could fade into the past, the fusion cue remained throughout. This may be why fusion ultimately dominated.

The question we have to address is whether we could imagine a case in which there were no cues at all, either for segregation or for fusion. That situation would seem to hold in a long burst of white noise (after the simultaneous onset cue for fusion faded away), since white noise, being the result of chance, would haphazardly contain fusion and segregation cues at random. One normally thinks of white noise as a coherent blast of sound. It is certainly not heard as a collection of separate sounds, each in its own frequency band. This would argue in favor of fusion as the default situation.

If we began by playing a narrow-band burst of noise, and then followed it by a wide-band spectrum that contained the frequencies of the first burst at the same intensities as in the narrow band, we would hear the narrow-band burst continue for a short time into the wide-band noise. But eventually we would hear only the wide-band noise. This implies that the default state is fusion.

There are also phenomena in music that point to the same conclusion but I shall defer discussing them until chapter 5.

Reallocation of Intensity and Timbre Information The observation from Helmholtz is relevant to another issue that we considered earlier. What happens to the properties of individual components when they give up their identities and become part of a mixture? Helmholtz's observation is valuable because he was able to observe the process as it slowly took place and to document the resulting changes in his experience. First, as regards loudness, he observed that when the high tone lost its separate identity, not only did it become weaker but the lower tone became stronger. Here is a puzzle: Why did the lower tone appear to get louder? After all, if the loudness of the high tone is to be allocated elsewhere, it should be to the global sound into which it it merging its identity. The lower sound is only another portion of the global sound. Why should it be given the loudness? I do not think that it was. I think this is a misinterpretation by Helmholtz of what he heard. It was the global sound that became louder, not the lower component. We have to remember that since the two sounds had the frequency ratio 2:1, they corresponded to the first two harmonics of a complex tone whose fundamental is equal to the frequency of the lower tone. Therefore this complex tone should have had a global pitch that was the same as the pitch of the lower tone (the fundamental) taken alone. For this reason it makes sense to believe that the lower-pitched tone that Helmholtz heard as getting louder was the global pitch of the now-unitary complex tone. This interpretation is supported by his report that the lower tone changed in quality as the fusion took place, which is what we would expect if the fusion of

high and low sounds led to a computation of global quality of a single spectrum containing this combination of components.

Helmholtz's observation of the strengthening of the lower tone, with a concomitant weakening of the upper one, resembles my early observations with Rudnicky on simultaneous sequences of tones. As the reader may recall, we used two tones of different frequencies, each having its own rate of repetition. At certain points in the sequence, the high and low tones started exactly at the same time. At such points of synchrony, we noticed, as Helmholtz did, that the higher tone became weaker and the lower one stronger.

The agreement of the two sets of observations supports the following conclusions: (1) When a high tone loses its separate identity, its energy is allocated to the complex tone whose pitch is determined by the fundamental. (2) The perception of the high tone as a separate entity is traded off against hearing its contribution to the global tone; the stronger one interpretation is, the weaker is the other.

The Consequences of Simultaneous/Sequential Competition

Although I have presented the organization of sequentially and simultaneously presented acoustic components in separate chapters, it has been evident that they interact. When we consider how this happens some new questions arise.

One of these concerns the role of timbre. Different ideas that I have introduced about how timbre related to organization seem contradictory. I have said that timbre emerges when a set of *simultaneous* components was partitioned into subsets by processes of scene analysis. An example of this would be the listener's ability, when listening to a voice accompanied by a guitar, to know that two different streams were present, each with its own characteristic timbre. However, I have also shown that similarities in timbre are used to group sounds sequentially. An example of this would be the ability to follow the melody of the voice, even when the guitar and the voice crossed each other in frequency.

Timbre, then, is both cause and effect of perceptual organization. This is not a contradictory position as long as it can be assumed that the timbre is always formed first by spectral segregation and then, after being formed, affects sequential integration.

However the interaction between simultaneous and sequential organization is not always so simple. Often we segregate a part of a mixture, A + B, because we have just heard A in isolation. Do we use the timbre of A, derived when we heard it alone, as a way of detecting its presence in the mixture? How could we argue that we do this if

we believe that the timbre of a subset of components is computed only after it has been segregated from the mixture?

Perhaps a way out of the dilemma is to distinguish between the *timbre* of a subset of acoustic components, which is a perceptual quality, and the *spectral composition* of that subset, which is a physical property. Then we can argue that when a sound, A, precedes a larger mixture, B, the thing that the auditory system tries to pull out of the mixture is not a timbre that matches A's (since timbre, a perceptual quality, has not yet been computed), but a combination of spectral components that matches A's. Then after the scene-analysis heuristics extract a subset of components, the timbre of that subset is computed. Timbre, according to this view, would be a perceptual description that was created after the parsing and grouping were done.

The timbre might be affected by the organization in even more complicated ways. For example, suppose that the presence of sound A1 before a more complex sound caused A2 to be extracted from the mixture. Perhaps A1 and A2 are not identical but are similar enough that A2 is extracted. It seems possible that the grouping of the components labeled A1 and A2 could affect the perceived timbre of both of them. If the auditory system thought that it was hearing only one long sound, the timbre assigned to A1 and A2 might be merged or averaged over time in some way. If A1 and A2 were truly parts of the same sound, this averaging would tend to correct errors in the calculation of the timbre of A2 that might result if it was not perfectly extracted from B.

The reader will have observed that the trick used to resolve the contradiction between timbre as cause and effect was to distinguish between physical properties and perceptual ones and to let the physical ones be the causes and the perceptual ones be the effects. This strategy leads us to the following question. Can a perceptual property such as timbre never be a cause of organization? Must we say that in sequential grouping, the auditory system looks always for a repetition of a physical pattern and never for a repetition of a perceptual effect? The only way to be able to decide this question would be to find different physical patterns that were heard as having the same timbre (the metameric timbres that we described in chapter 2). If two spectra, despite having metameric timbres, had different ways of grouping with other spectra, then it would be the physical pattern and not the timbre that controlled the grouping. If, on the other hand, patterns that had the same perceptual timbre always grouped in the same way with other sounds, we might have reason to conclude that the timbre itself was controlling the grouping.

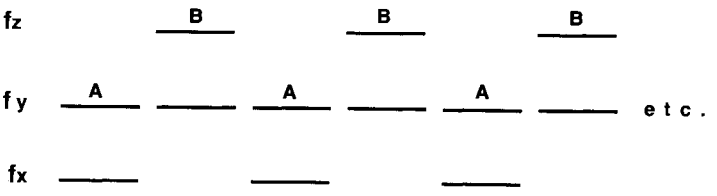


Figure 3.19
"Ternus" effect in audition.

Even in this case it might be whatever physical property caused the similarity in timbre, and not the timbre itself, that controlled grouping. The only indisputable way I can think of to prove that timbre, and not its physical causes, causes a grouping, would be to alter a perceived timbre through some manipulation of its prior context (but without changing its own spectral content), and then to show that it was the timbre, as altered, that governed its grouping with subsequent events.

We know that matches between the components in current and recent spectra can cause those components to form sequential groups and to be detached from the larger spectra of which they are a part. We saw this in the ABC experiment of Bregman and Pinker, where a sequential grouping AB was able to destroy the integrity of a complex tone BC. One would think that there should frequently be accidents in everyday listening in which components of two successive spectra, X and Y, matched one another by chance and were heard as separate sounds. Why do we never encounter such decompositions in everyday listening? Normal sounds seem to hold together.

One inhibiting factor in many cases would be the harmonic relations among the partials within X and within Y. These would oppose the sequential grouping. So would all the other factors that have been shown to affect spectral fusion. But even harmonicity might not always oppose inappropriate sequential integrations. Here is a laboratory example. Suppose we created an alternation of two tones, A and B, each formed of two harmonically related partials, and arranged it so that the bottom partial of the first one matched up with the top partial of the second one. Let us suppose that the frequencies were fx, fy, and fz as shown in figure 3.19. Let us also assume that the ratio between the frequencies fx and fy is the same as between fy and fz. This being the case, we can think of note B as an upward transposition of note A. Nevertheless, if fx, fy, and fz are far enough apart in frequency and the sequence is fast enough, the successive repetitions of the partials at fy will form their own stream. We would hear a

continuous repetition of fy, accompanied by two additional streams, one consisting of fx and the other of fz. Here it is the partials rather than the complex tones as a whole that act as units. How is it that the constant harmonic structure does not fuse the partials of tones A and B so that we can hear a complex tone alternating between two pitches?

The reader who is familiar with the field of visual perception will have noticed the resemblance between this example and the Ternus effect. Think of the figure as describing the vertical positions of two dots as they change over time rather than the frequencies of two partials. In vision, the dots will sometimes be seen to move as a pair (the Ternus effect) or the middle one will remain stationary and the other one will alternate between a high position and a low one. One of the main determinants of the percept is whether the successive displays are separated by a time gap. If they are not, and the middle dot remains on continuously, it will tend to remain stationary while the other dot skips back and forth around it. With the time gap, the pair appears to move as a whole.[433] Apparently the continuous presence of the dot at the middle position is treated as an overwhelming cue that it has not moved. It is unlikely that it has been replaced by an accidentally corresponding part of a different object.

The factor of continuity is important in our auditory example as well. The interpretation of the partial at fy as a separate sound is made stronger if there are no silences between tones A and B (that is, where fy stays on all the time). Apparently the auditory system treats it as vanishingly unlikely that a harmonic of one sound would be exactly replaced by some other harmonic of a second one with no discontinuity.

Even with gaps, however, our laboratory example is sometimes not heard as two coherent sounds. Yet if we do not always hear A and B as coherent tones how can we hear a descending musical interval (say C-sharp followed by B-flat) as moving downward in an ordinary piece of music? If each harmonic of the first tone is searching for the nearest harmonic in the second to group with, there will be many cases in which this process will pair a partial of the first with a higher one in the second. It appears that instead of this piecemeal approach, the tracking of the note's movement is operating on the whole spectrum. The auditory system seems to be apprehending the motion of a global mass of sound upward or downward. In fact, as we shall see in chapter 5, a musical tone acts so much as a unit that we can usefully consider it to be describable by its fundamental frequency (let us just call it pitch). [This simplification allows us to show that there are parallels in music between the sequential and simultaneous organiza-

tions of pitches and the pure-tone grouping and segregation that we have observed in the experiments of the earlier chapters.] Why do the musical tones always function as units? There may be more than one answer to this question.

One is that all the factors that we have described as facilitating spectral fusion (correlated AM, FM, and so on) would be present internally in each tone, binding it together. Among these binding factors we might list spectral density. We should note the laboratory example that I have just described (together with many other products of the laboratory) is a spectrally sparse signal with few harmonics. We know that partials that are relatively isolated in the spectrum are easier to capture.[434] Therefore the spectral density of natural signals such as instrumental tones may prevent their being torn apart by sequential grouping of their partials. In sparse signals the individual harmonics may be able to act more independently.

Another factor that may contribute to the internal stability of musical tones is the fact that the formation of a sequential stream in a particular frequency region takes some time to build up in strength. An accidental coincidence of a subset of harmonics in two successive musical tones of different pitches is likely to be a brief event. Unless it occurs over and over again, the coincidence is likely to be discarded by the auditory system. It is probably this reluctance to form streams that prevents the helter-skelter appearance and disappearance of streams based on momentary matching of partials. In the experiments in which partials did group sequentially to break up complex tones, the alternation of the partial with the complex tone was repeated over and over to establish a strong sequential stream.

One requirement for a simple explanation of anything is that the explanation be free of contradiction. Some years ago, I argued that perception was a description-forming system that provided explanations of the mixture of evidence that reached the senses, and that one of its properties was the tendency to avoid contradictions.[435] I pointed out that consistency cannot be defined in its own right, but depended on higher order rules that controlled the kinds of descriptions that the system would allow to coexist with one another. An example of a consistency rule is that a region A in space cannot be both in front of and behind another region B at the same time. Different domains of experience were seen as having different types of consistency requirements.

Here is an example of an effect of the simplicity principle in auditory organization. It seems to embody a rule of consistency. The example is based on an experiment by Yves Tougas and myself at McGill University.[436] The pattern of tones that was used is shown in

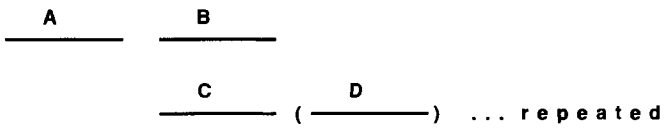


Figure 3.20
Competition of groupings. Tone D is presented only in some conditions.

figure 3.20. It was like the one used by Bregman and Pinker in that a tone, A, was alternated with a mixture of two tones, B and C. A was at the same frequency as B and was intended to capture it into a sequential stream, AB–AB–AB. . . . The only difference was that in some conditions there was a fourth tone, D, at the same frequency of C and following it in the cycle. In the conditions that had no fourth tone, a silence equal in length to tone D was added to the cycle in its place to ensure that the cycle always lasted for the same length of time. The pattern was played repeatedly to listeners who were asked to judge how clearly A and B could be heard as a repeating pair. They heard this AB grouping more clearly when D was present.

We explained these results by the following argument. D tends to capture C into a CD stream. This competes with the spectral fusion between B and C and weakens it. Due to this weakening, B is freer to group with A. We can use a metaphor to clarify this reasoning. Instead of tones, let A, B, C, and D, be Alice, Bob, Carol and Don, respectively. Bob and Carol are going together, but Alice is interested in Bob. Cleverly, she introduces Don to Carol. When Carol becomes attracted to Don, the bond between her and Bob decreases and Alice is more successful at persuading Bob of her attractiveness.

The results follow from the fact that spectral fusion competes with sequential grouping. The auditory system acts as if it is implementing the rule that a sound cannot exist in its own right with its own properties and at the same time give up its own properties to be fused into a larger organization. This consistency requirement can also be viewed as an example of the rule of belongingness or exclusive allocation that we discussed earlier. Later we will examine evidence that suggests that this rule is limited in two ways: First it is not all or nothing in nature, but quantitative. For example, when the grouping of B with A increases to some extent, B's grouping with C decreases to some extent. Second, the rule of exclusive allocation will not be followed when a component of sound fits in very well to two competing organizations (that is, when there are several cues favoring its inclusion in each organization). Sometimes instead of being appor-

tioned to the two organizations, it is duplexed to the two. That is, it seems to be given almost completely to both.[437]

To the extent that the exclusive allocation rule holds, constraints on grouping can propagate to other frequencies and other times. The Gestalt psychologists would have called the frequency-by-time representation of sound a field and would have argued that it existed as a real field in the brain. The definition of a brain field is any analog of a set of relations in the outside world that acts in the way that a field does in physics. In a physical field, effects propagate from one part to another by crossing the intervening space. A local property of a field cannot affect something far away in the field without influencing things that are closer to it. Influences in a field are also usually stronger between elements that are closer together in the field. The pictures that I have constantly been drawing, showing time and frequency as the two axes, would, for the Gestaltists, be portrayals of actual fields in the brain. For them this would explain how constraints could propagate from one part to another of the frequency-by-time representation of our sounds, so as to exert their influences in favor of a consistent or simple organization.

There is an analogous idea in machine vision. In understanding a drawing, for example, a computer program could be trying to establish a label, such as "convex corner," for a particular junction of lines in the drawing.[438] Yet the information in the immediate vicinity of the junction (the local information) might be ambiguous and support a number of alternative labels. One approach to resolving ambiguity is to first create a description of the drawing in which we assign to each junction all the labels that the local evidence will support, and subsequently to compare pairs of adjacent junctions to see if their labels are consistent with each other.[439] When such comparisons are made, the number of possibilities for valid labels drops dramatically. This procedure will propagate constraints across the description, a decision about one line junction acting to force decisions about neighboring junctions. A generalized theory based on this method has been called "relaxation labeling."[440] Each local region of a pattern is assigned a set of labels. Associated with each label is a strength. An algorithm that allows labels to interact leads to the raising of the strengths of some labels and the reduction of the strengths of others.

Properties of relaxation labeling include the following:

1. Mutual support relations exist in the labeling of neighboring elements. In the foregoing auditory example, this type of relation might be in the form of a rule that says that if one tone is strongly attracted to a second, then the second should be strong-

ly attracted to the first. In our example, "if C goes with B, then B goes with C."

2. Mutual antagonism relations exist among certain of the labels associated with a particular element. Hence in our auditory example, this would say that the following two labels on B would tend to interfere with one another:

—"B belongs with A in a stream whose base pitch is the pitch of A";

—"B belongs with C in a stream whose base pitch is the pitch of C."

I am assuming that because there are incompatible properties of the streams to which B is assigned by these two labels, if one label is asserted strongly this will weaken the other.

3. Certain physical relations between tones also affect the strength of labels. These are the properties that we have been calling heuristics of scene analysis. In our example, A will be attracted to B and vice versa by frequency proximity. So will C and D. B will be attracted to C by synchronous onsets and offsets.

In view of these rules, when D is added to the pattern the CD attraction (reflected both in the labels of C and D in a mutually supportive way) will enter the situation. By rule 2 (antagonism of labels) this will weaken the attraction of C toward B. By rule 1 (mutual support of labels) this will reduce the attraction of B toward C. By rule 2 (antagonism) this will strengthen the attraction of B toward A. By rule 1 (support) this will increase the attraction of A toward B. The perceptual effect of the last step is to favor the perception of A and B as a stream, which is what we find.

Admittedly sketchy, the above outline merely traces out a path of relationships whereby the presence or absence of D in the pattern could make a difference in the perception of A and B as a stream, and maps it informally onto a relaxation labeling system.

The experiment is interesting because its results are hard to explain by a filter theory of attention.[441] In such a theory, the formation of a stream in a particular frequency region is due to the tuning of a filter to that frequency. In our example, when A is heard in isolation, this tunes a filter to that frequency and when this filter subsequently encounters the BC mixture, it strips out B. The output of this filter then contains A and B, which are then perceived together as a pair. In a similar way, a second filter would become tuned to D's frequency by encountering it in isolation and, on the next cycle of the pattern, would act to strip C out of the BC mixture. This theory ought to find

no reason why the high- and low-tuned filters should not act quite independently of one another, any later recognition process selecting the output of either one depending on its needs. Such a view would expect that when we are listening to the high stream, we are attending to the output signal of the filter that selects A and B and that the presence or absence of D in the low stream should not affect the clarity of this signal.

Although the effects of D are not consistent with a simple filter theory, the relaxation labeling model with which we have contrasted it is only one possible embodiment of a larger class of theories: field theories, in which large numbers of elements interact to determine a perceptual result and where consistency of interpretation is enforced in some way. Other embodiments of field theory might serve just as well.

The question of consistency of interpretation is related to the question of what happens to the residual when one portion of a mixture is extracted by processes of organization. If we start with a mixture PQ, and we extract P, what happens to Q? If P is now separate from Q, then Q should also be separate from P. Certain perceptual results should occur. The first is that Q should become freer to group on its own with other sounds. We saw that in the previous example. The second should be that it is heard as having the properties that are left behind when the properties of P are removed.

We can illustrate this second principle with a signal in which there is a repeated alternation of two sounds, A and B. A is a noise burst that is limited to the narrow band between 100 and 900 Hz. B is a wider band of noise limited to the band between 100 and 1,700 Hz. Acoustically, therefore, B includes A as a component band. I have tried to make a visual analog of this in figure 3.21, where the vertical dimension represents frequency. The A bursts are three units high and the B bursts are six units high. They are shown separately on the right. When A and B are alternated without pauses between them, instead of hearing an alternation of two bursts with different properties, we hear A as continuously present, with a higher pitched noise burst periodically accompanying it. The explanation is that the part of B that is identical with A tends to separate from the rest of B and group with A. This leaves the residual higher pitched part of B to be heard as a separate sound. If the pause between A and B is long enough, B will not be partitioned into A plus a residual.

The formation of separately audible residuals is an important part of the primitive scene analysis process, a property which, as we shall see in chapter 4, distinguishes it from schema-governed segregation.

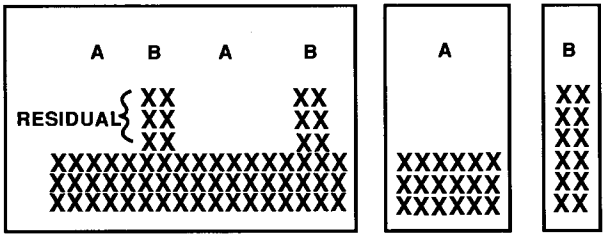


Figure 3.21
Formation of a residual.

Apparent Continuity and Contralateral Induction

We have seen that portions of a spectrum of sound can be separately allocated to different streams. One of the factors that led to the segregation of a particular part of the spectrum was the fact that a preceding sound was able to capture that part into a sequential stream. In this section, I would like to consider two phenomena that are examples of the same factor in operation but, because they are so striking, have been given special attention. They are both illusions, but the purpose of this section will be to show that they occur because basic principles of scene analysis are fooled by a sound pattern that would seldom occur in natural listening situations. Research on the two illusions is surveyed in a book by Richard Warren.[442]

One of them is a phenomenon that we looked at briefly in chapter 1. We can call it "the continuity illusion," although it has been given many other names. It is the illusion that one sound has continued behind a louder interrupting sound even when the softer sound is not really there during the interruption. A case of this, created by Gary Dannenbring, was illustrated in figure 1.15 of chapter 1.[443] In that example, a pure tone glided up and down repeatedly in frequency, but at a certain point in each glide, part of the tone was replaced by a silent gap. When that was done, the glide pattern was heard as having gaps in it. However, if the silent gaps were replaced by noise bursts, the gliding tone was heard as continuing through the noise and as not having any gaps. We saw that the effect was an auditory example of the perceptual closure effects that had been investigated by the Gestalt psychologists.

The second effect has been mainly studied by Richard Warren, using a consistent term, so I will use his name for it, "contralateral induction."[444] Here is an example. If, using headphones, we play a tone to the left ear, it will be heard on the left. However, if we play a loud noise burst to the right ear at the same time, the tone will be

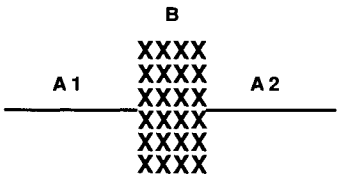


Figure 3.22
Labeling of the parts of the continuity illusion. A1 and A2 are parts of a longer tone A. B is a louder interrupting sound.

heard as closer to the midline of the listener's body. The contralateral noise will have induced an erroneous perception of the location of the tone.

I have chosen to put the two phenomena together, because both are cases in which spectral components are taken from one signal, A, and reallocated to a second signal, B, because they are assumed by the scene-analysis processes to be part of B. In the case of our example of the continuity illusion, a part of the spectrum of the noise has been allocated to the stream containing the tonal glide, supplying sensory evidence for the glide so that the latter is not perceived as having any gaps in it. We will have to explain both why the auditory system has detected a need for a subdivision of the noise spectrum and how it knows what to take out.

Our example of contralateral induction can be explained by assuming that a reallocation of spectral components has also taken place. In this case, scene analysis has assigned part of the inducing noise that arrives in the right ear to the same stream as the energy from the tone. Because of this, the tone is interpreted as having provided energy to both ears. Therefore it is heard as being located nearer to the midline of the body than it would have been if the contralateral noise had not been there. Again we have to explain how the system detects the necessity to reassign energy and how much of it to reassign.

The Continuity Illusion

I would like to begin with a discussion of the continuity illusion. The illusion has been investigated in many forms, but in every case, the experimenter deletes parts of a softer signal and replaces them with a louder sound, and the listener hears the softer sound as continuing unbroken behind the louder one. For our discussion, let us use the following terminology, illustrated in figure 3.22. The horizontal line represents an originally continuous sound whose center section has

been deleted and replaced by a short, loud sound. The block of X's represents that interrupting sound. The continuous sound as a whole is called A and the interrupting sound, B. The part of A before the interruption is called A1 and the part after it is called A2.

Richard Warren has described and classified the known examples.[445] Perhaps the simplest case occurs when a soft noise burst seems to continue behind a louder burst that has an identical spectrum. In the most complex case, a spoken sentence appears to continue through a loud noise burst. This latter instance has been called "phonemic restoration" because when the noise burst replaces a phoneme in the speech, the brain of the listener will often restore the percept of a phoneme so that the partially deleted word seems complete.

The continuity illusion can be explained from the point of view of Gestalt psychology as an example of the general tendency for discontinuous perceptual events to exhibit closure if their properties match one another.[446] However I agree with Warren, who has done a great deal of insightful research on the continuity illusion, that we can understand it better by seeing it as a perceptual compensation for masking. In a natural listening environment when a sound is momentarily interrupted by a much louder one, the louder sound could totally mask the sensory information arriving from the weaker one. In such circumstances, it would be incorrect for the auditory system to interpret the weaker signal as having stopped during the interruption, especially if the weaker signal becomes audible again after the interruption. In Warren's view, the auditory system improves perception by restoring what it can infer as having been masked by the signal. While I do not think that the auditory system need know anything specifically about masking to achieve this task, I find Warren's view stimulating.

The auditory system, however, can never get something for nothing. If evidence is missing, you can never get more information by inferring what might have been there. According to a scene-analysis approach, the way that the auditory system improves the quality of its decision making in the case of an interrupting sound is by avoiding being fooled by the evidence that *is* there. For example, let us refer back to figure 3.22. At the transition between A1 and B, there is a sudden spectral change, a boundary, at the point where the interrupting signal starts. After this point, no specific evidence for the existence of A is present. Yet if this boundary were interpreted as the offset of A, the auditory perceptual process would incorrectly estimate the duration of A and describe A as a signal that had only those features that existed prior to the interruption. As a result, when

processes of pattern recognition were applied to the signal, they would try to recognize a sound that had only those properties that occurred prior to the interruption and would not integrate those properties that occurred after the interruption as part of the description of the same sound. By interpreting A1 and A2 as two separate signals, the pattern-recognition process would come up with the wrong interpretation of the signal. We can get an idea of what such mistakes might look like from the following visual example of a printed sentence.

They should be* in the performance.

If the asterisk in the above sentence represented an inkblot that covered that position in the sentence, and if the absence of a visible letter at the position of the asterisk were taken as equivalent to a space, the reader would read the sentence as "They should be in the performance." But if the asterisk were treated as missing data, then the listener might go into his mental dictionary looking for a word of about five letters, starting with "be" and ending with "in", and find the word "begin". Similarly, in visual examples involving the recognition of the shapes of objects, when the outline of a shape nearer to our eye interrupts our view of a more distant object, the visual system will come up with two different descriptions of the distant object depending on whether it incorporates the shape of the occluding edge into the description of the more distant object or leaves it out. This problem was discussed in chapter 1 and an example was shown in figure 1.5.

The perceptual restoration of the word "begin" in the previous example would be correct if the word had actually been continuous through the inkblot. We should observe that the correctness of the interpretation of the sensory evidence that occurs on both sides of an interruption as evidence for a single object (or event) depends strongly on the correctness of the assumption that our view of the target has been occluded or interrupted. For this reason, the perceptual systems, both visual and auditory, must use a very accurate analysis of the structure of the sensory evidence to determine whether the parts separated by the occluding material show sufficient agreement with one another to be considered parts of the same thing or event.

As far as the continuity illusion is concerned, then, I would not only want to see it as a compensation for masking but would like to show that it is a part of scene analysis, and that the perceptual process that is responsible for it uses many of the same rules for scene analysis that we have already discussed.

I would like to start by proposing a set of principles that I see as governing the continuity illusion and then go on to illustrate them through examples taken from the research literature.

The account can be broken down into two parts. When the illusion occurs, the listener's brain must come up with answers to two questions, a "whether" question and a "what" question. In answering the "whether" question, it must decide whether the softer signal has stopped or continued at the point in time that the louder signal occurred. In this section we will see that primitive scene analysis makes a substantial contribution to this decision. This contribution can be made without recourse to learned knowledge that is concerned with the structure of the softer signal. By "learned knowledge," I am referring to such things as rules concerning the nature of the speech signal or the inventory of words in the language. The "whether" question is presumed to be answerable largely on the basis of purely acoustic cues that indicate that the softer signal has not ended but has merely been interrupted by another sound.

When the brain has determined that the softer signal did continue on behind the louder one, it must go on to answer the "what" question: What are the properties of the softer sound during the period when it was inaudible? For example, when the softer sound is a rising frequency glide of a pure tone both before and after the louder sound, and the two audible parts fit nicely together as parts of a longer glide, the brain concludes that the missing part was the middle portion of the long glide. As a second example, when the softer sound is a spoken sentence, it concludes that the missing part was the speech sound that is predictable from the parts of the sentence that were clearly audible. This question of what to restore is clearly based on a process of prediction that is able to use the audible portions to infer the nature of the missing parts. In the case of the ascending glide, the process must be able to interpolate along some dimension such as frequency. In the case of speech sounds, the restoration is clearly based on knowledge of the words of the language.

It is important not to confuse the two questions. I will argue that the "whether" question falls within the scope of scene analysis as we have described it previously. In deciding that A1 and A2 are parts of the same signal, the scene-analysis process is doing its job of assigning parts of the sensory input to a larger group that probably has all come from the same source. In doing so it alters the nature of the recognition process so that a sound that starts with A1 and ends with A2 can be recognized. Primary stream segregation is concerned with acoustic properties whose meaning is not yet interpreted. For this reason it cannot determine the specific nature of the missing material. The

latter must be determined by a more complex process of pattern recognition. Therefore answering the "what" question depends on processes of a wholly different nature, which are outside the scope of this volume. We are employing a dichotomy here, between primitive and schema-based scene-analysis processes, that we will discuss in more detail in chapter 4. We are entitled to believe in a two-component process if we can show that the two components do different jobs and base their activity on different sorts of information.

Rules Governing the Generative Process

Let me begin by briefly listing the fundamental rules that govern the process of deciding whether a sound A has continued through an interruption by another sound B.

- There should be no evidence that B is actually covering up a silence between A1 and A2 rather than the continuation of A. This means that there should be no evidence that A actually shuts off when B starts or turns on again when B finishes.
- During B, some of the neural activity in the auditory system should be indistinguishable from activity that would have occurred if A had actually continued.
- There should be evidence that A1 and A2 actually came from the same source. This means that the rules for sequential grouping would normally put them into the same stream even if they had been separated by a silence instead of by B.
- The transition from A to B and back again should not be capable of being interpreted as A transforming gradually into B and then back again. If it is, the listener should not hear two sounds, one continuing behind the other, but simply one sound, changing from one form into another and then back again. The criterion for hearing "transformation" rather than "interruption" is probably whether a continuous change can or cannot be tracked by the auditory system.

In short, all the sensory evidence should point to the fact that A actually continued during the interruption and has not either gone off altogether or turned into B. Let us see how this works out in detail.

The "No Discontinuity in A" Rule

The first of the preceding rules said that there should be no reason to believe that there was not actually a silence between A1 and A2. We can refer to this as the "no discontinuity in A" rule. One source of evidence that is relevant to this decision is what happens at the bound-

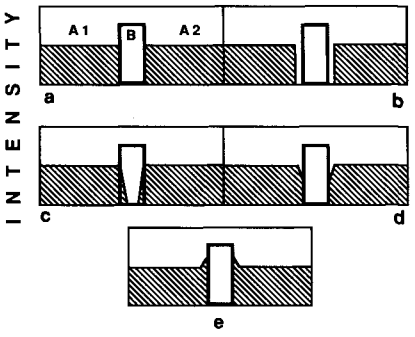


Figure 3.23
The effect of the boundary between sounds on the continuity illusion. (a) B is much louder than A and in the same frequency region. It masks the offset and subsequent onset of A. (b) There is a silence between A and B so that the offset and onset of B are audible. (c) B is not loud enough to mask the offset and subsequent onset of A. (d) A is perceptibly dropping in intensity just at the onset of B and rising just after B's offset. (e) A is perceptibly rising in intensity just before B and dropping just after B.

ary between A1 and B and between B and A2. It has been shown that if there are short 50-msec silent gaps between B and the two parts of A, then A will not be heard as continuing behind B.[447] This condition is diagrammed in part b of figure 3.23. I think the lack of perceived continuity occurs because the auditory system can hear A1 turning off before B comes on, and then, after B, can hear A2 turning on again. Even if the amplitude of B is sufficient to mask the offset of A1 and the onset of A2 its temporal separation from these two events prevents it from doing so. The idea that continuity occurs only when the auditory system cannot detect a drop in intensity when A turns off, or a rise when it turns back on again, is the essence of several theories of the continuity illusion.[448]

It is not necessary for the auditory system to detect a complete offset and onset before it becomes unwilling to hear A as having continued through the interruption. Gary Dannenbring and I did an experiment that did not insert silences between A1, B, and A2. Instead we introduced a 10-dB drop in intensity of A1 just before B occurred (matched by a 10-dB rise just as it reappeared after B). We found that the auditory system was less prepared to hear A continuing though B than if no such drop (and rise) had occurred.[449] This condition is shown in part d of figure 3.23. The utility of such a tendency should be obvious. If sound source A in the world is in the process of shutting off just as B appears, it should not be heard as continuing

through the mixture. However, another result in the same experiment surprised us. In one condition, A suddenly began to increase in intensity just before B and then decreased in intensity upon its reappearance after B, as in part *e* of the figure. In this condition, too, there was a reluctance to hear A continuing during B. It seemed that any discontinuity in A, near the A-B boundary, interfered with the perceptual continuation of A, and that this discontinuity in A was interfering with the allocation of the A-B boundary to B.

Let me explain what I mean. In the case where A remains steady up to the A-B boundary, as shown in part *a* of figure 3.23, the spectral discontinuity at that boundary is attributed to B, not to A. It is as if the auditory system were saying (in the steady case) that the discontinuity marks the beginning of B, but has no relevance for the description of A. It resembles the exclusive allocation of properties discussed in chapter 1. We saw that in a drawing, a line might be interpreted as portraying the shape of the outline of a nearer object where it occludes a farther object from view. In such a case, the line would not contribute to the description of the shape of the farther object. Perhaps the temporal boundary between A and B is treated in a similar way; it is assigned to B and helps delimit its duration, but not A's. We still have to explain why sometimes it is assigned only to B, and why under other circumstances it is assigned to A as well. Perhaps, as a general rule, it tends to get assigned to the louder sound. However, if there is some indication that the softer sound is starting to undergo a change, the discontinuity can get assigned to the softer one as well and we will not hear it as continuing behind the louder.

Another example of the fact that boundaries tend to be assigned to the interrupting (louder) sound occurs in a demonstration created by Ranier Plomp.[450] A 1,000-Hz tone is frequency-modulated at 15 Hz. After every three modulations there is a white noise burst or a silence replacing the signal. In the case of a silence, the listener hears a series of triplets of warbles, as shown in part 1 of figure 3.24. In the case of the noise, shown in part 2, there is no experience of triplets. The sound is simply heard as a continuous sequence of modulations that is periodically interrupted by a noise burst. This means that the mere absence of the FM warbles was not sufficient to define the groupings that led to the perception of triplets. The triplets had to be separated by silence. The beginning and ending of an interrupting sound will define its own temporal boundaries but not the boundaries of the thing it interrupts.

The ability of the auditory system to detect a discontinuity in A may be involved in another fact about the continuity illusion. Suppose we start with a very weak A and a B of medium loudness and

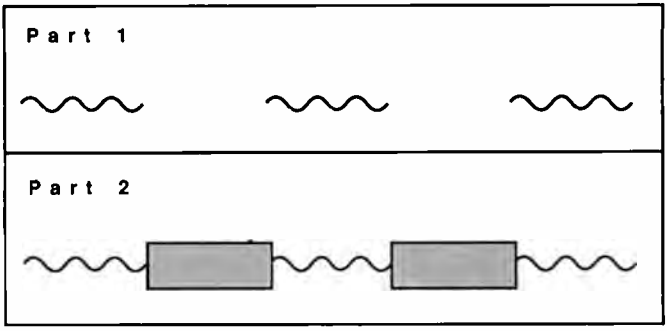


Figure 3.24
Part 1: bursts of a tone, modulated in frequency. Part 2: the same tone bursts alternating with noise bursts. (From Plomp 1982.)

alternate them repeatedly with no gaps. At these loudness settings, A will sound continuous. Suppose, then, that as the alternation proceeds we gradually increase the loudness of A. At a certain point, A will no longer sound continuous but will seem to be pulsing in loudness, dropping at the point where B turns on, and rising again after each B. This point has been called the pulsation threshold. (Obviously, when A becomes much louder than B the listener will hear A turning on and off repeatedly.)

The pulsation threshold has been interpreted in the following way: "When a tone and a stimulus S are alternated (alternation cycles about 4 Hz), the tone is perceived as being continuous when the transition from S to tone causes no perceptible increase of nervous activity in any frequency region."[451] In other words, the frequency receptors responding to A are allowed to be more strongly stimulated during the interruption, but not more weakly stimulated. The reason for this requirement is obvious if the auditory system is to infer that A did not turn off during the interruption.

We also know that the continuity illusion becomes less compelling when the weaker sound is presented to one ear and the more intense sound to the other.[452] This would also be expected if the louder sound is required to mask the offsets and onsets of the weaker before continuity may be heard. Sounds are less effective in masking others if they are presented to different ears.

The "Sufficiency of Evidence" Rule

The second rule said that during B, some of the neural activity should be indistinguishable from activity that would have occurred if A had actually been continued. This rule can be called the "sufficiency of

evidence" rule. Warren and his colleagues have described it in this way: "If there is contextual evidence that that a sound may be present at a given time, and if the peripheral units [of the auditory system] stimulated by a louder sound include those which would be stimulated by the anticipated fainter sound, then the fainter sound may be heard as present."[453] This description can make sense of a number of effects that have been seen in experiments.

First, as we have already seen, if the interrupting sound is not sufficiently louder than the softer one, the softer may not appear to continue underneath the louder.

A second observation was made by Warren and his colleagues. As part of Warren's project to show that the continuity illusion was a compensation for masking, an experiment was done that compared two sets of conditions: those in which a louder sound would mask a softer one, and those in which the two sounds could be used to generate the continuity illusion.[454] Both the softer (A) and the louder (B) sounds were 300-msec sinusoidal tones. The authors expected that the frequency relations that allowed B to mask A would be the very ones that favored the perceptual continuity of A behind B. In the assessment of the continuity illusion, A and B repeatedly alternated with one another with no gaps. The listeners were told to adjust the loudness of A to the loudest level at which they could still hear it as continuous (the pulsation threshold).

The frequency of the louder tone was always 1,000 Hz, but the frequency of the softer one was varied in different tests for the pulsation threshold. The reason for this is that the masking of one pure tone by another depends on the frequency relations between them, and the experiment tried to determine whether this would also be found for the continuity illusion. The masking of the softer tone by the louder was also directly measured. In the masking test, the louder 1,000-Hz tone was kept on all the time while the weaker tone was repeatedly turned on and off. The masking threshold was the faintest intensity of A at which it could be heard going on and off. Again the frequency of tone A was varied on different tests. The results showed that the thresholds in both the continuity tests and the masking tests followed the same pattern. Let me describe it first for the masking. The masking by B was greatest when A had the same frequency and fell off with increasing frequency difference in either direction. However, this reduction in masking was not symmetrical, with A suffering more when it was higher than B in frequency. This is a typical finding and is referred to by the expression "upward spread of masking." The masking effects of a tone are greater upon frequencies higher than it than upon lower frequencies.

The results for illusory continuity were similar. Continuity was best when the frequency of A and B was the same, and declined (required A to be softer) as A's frequency deviated from B's. Furthermore, the same asymmetry in this effect was observed for continuity as for masking. When A was higher in frequency than B it could be made louder than it could be made when it was equally far below B in frequency and still be heard as continuous.

According to Warren's view the similarity between the two sets of results is due to the fact that continuity is a compensation for masking. The auditory system will restore the signal under just those conditions where, if it had really been there, it would have been masked. If this view were to be taken as literally correct, the auditory system would have to be equipped with a knowledge of which sounds can mask which other ones. As an indication of the complexity of the knowledge that it would have to have, consider another experiment that was done in the same series. Instead of using a tone as B, a broadband noise burst was used. This noise burst was filtered so that frequencies near 1,000 Hz were eliminated. Therefore it had a reduced masking effect when tone A was at 1,000 Hz. Correspondingly, it did not allow tone A to be heard as continuous unless it was very soft (and therefore capable of being masked by B). This would seem to point to a profound knowledge by the auditory system about the masking effects of a vast array of spectral patterns on one another.

There is, however, an explanation that does not require the auditory system to have such deep knowledge. This account says that rather than depending on a knowledge about masking, the continuity illusion depends in a very direct way on masking itself. Earlier we saw that if the auditory system detects onsets and offsets of energy in A at the points in time at which it abuts against B, it will not project A through the interruption. We can understand the close relation between continuity and masking by realizing that if these onsets and offsets are masked, the system will have no reason to infer that A has stopped and started again, and will therefore assume that A has continued through the mixture. The onsets and offsets of energy in a particular frequency region will be most effectively masked by a masker of the same frequency, and also more effectively by maskers below it in frequency than by those above it. The same idea will explain why a louder B is more effective in inducing perceptual continuity: It more effectively masks the offset and onset of A.

The idea that the discontinuities in A have to be masked also explains another observation. If A and B are tones, and there are gaps of silence between A1, B, and A2 (as in part b of figure 3.23) then the longer the gaps, the lower we must set the intensity of A in order for

it to be heard as continuing through B.[455] This is consistent with what we know about the masking of a sound by another one that either precedes or follows it (backward and forward masking). The following two facts are relevant: (1) the nearer in time a target is to a masker, the more it is masked, and (2) the more intense the masker relative to the target, the greater the masking.[456]

It is interesting to remind ourselves of the details of the procedure that the experimenters employed to measure masking in the experiments that showed the connection between masking and continuity. B was the masker. It remained at a constant amplitude. It was accompanied by a less intense tone, A, that was periodically pulsed on and off. The main clue that a listener uses in deciding whether A is present in such a pattern is whether any periodic changes (discontinuities) can be heard in a particular frequency region. This procedure, therefore, was probably measuring the masking of these discontinuities. Therefore the correspondence between the results in the masking and continuity tests are consistent with the idea that the masking of the onsets and offsets of A are a necessary prerequisite for the continuity illusion.

This is not the only prerequisite. What we hear during B must be consistent with what we would hear if A were still present but mixed with another sound. Therefore the role of the spectral content of B is twofold. On the one hand it must be of a proper frequency and intensity to mask any discontinuities of A at the A-B boundaries, and on the other hand, it must be such as to sustain the hypothesis that A is indeed present in a larger mixture. That is, the peripheral neural activity that signals the presence of A when it is alone must be part of the neural activity during B.

We can easily see why the masking of the onsets and offsets of A is not sufficient. Suppose that A was a pure tone and B a white noise burst with a silent gap embedded in it. The onset of B would mask the offset of A1. However, if B lasted only briefly after masking A1, then went silent, then reappeared just in time to mask the onset of A2, we would surely not continue to hear A through the silence. Otherwise we would never be able to hear silences. This is the reasoning behind the second rule in our earlier list, which said that the neural activity during B had to be consistent with the case in which A was actually present.

Notice that the requirement applies to neural activity not to physical acoustic energy. The reason for making this distinction is that there may be cases in which there is no actual acoustic energy in B occurring at the same frequency as the energy in A and yet we will still hear the continuity. This will occur when the neural activity normally

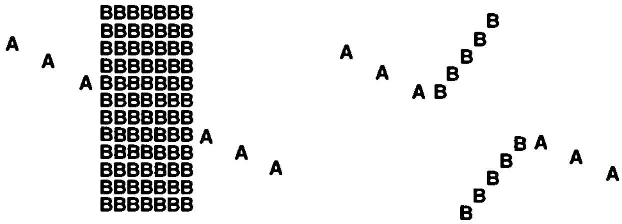


Figure 3.25
Illustration of the principle of simplicity in the selection of energy from an interrupting sound to complete the signal that occurs before and after it.

associated with A is stimulated by B. B might have the wrong frequency but be very intense. We know that a strong stimulus leads to activity in neural pathways far removed from those that are primarily tuned to the arriving frequency. Therefore, this spreading activation might activate the pathways that normally respond to A.

One of the most interesting things about the continuity illusion is that the important thing is not the match of the neural activity during B to the activity present during A1 and A2, but the match of this activity to the activity that *would have occurred* if A had really continued behind B. B must provide the stimulation that would have been provided by that missing part. For a perhaps ludicrously simple example of this requirement, let us suppose that there are two frequency glides that are aligned on a common trajectory, as shown on the left of figure 3.25. The discrete A's in the figure are to be taken as signifying a continuous pure tone that descends along the trajectory that they outline. The B's are meant to show a burst of noise that briefly interrupts the glide. The noise is shown as containing frequencies that match the parts of the glide that occur before and after it as well as the part that would occur underneath it. The diagram shown on the right contains the same parts of A, but most of B has been erased. Only two parts remain, one that continues the first part of the glide (but reversed in slope) and one that connects up with the second. Those parts of the noise burst that remain correspond to a sound in which B consists of a pair of parallel ascending tonal glides. If we listened to the sound shown on the left, we would hear, if B was loud enough, a long glide passing behind B. Such effects have been observed by Gary Dannenbring and Valter Ciocca in my laboratory.[457] However, if we listened to the sound shown on the right, we would hear a high-pitched glide first descending and then ascending again, accompanied by a low-pitched glide that first ascends and then descends.[458]

Why is this not heard even in the example shown at the left? After all, the necessary frequency components are there! It seems that although the percept is obliged to use the sound that is present in the B interval, it does not select from it in arbitrary ways. The selected part of B seems not only required to be consistent with the audible parts of the interrupted sound, but to be the simplest sound that is consistent with those parts. The auditory system can apparently select a subset of the available stimulation to use for its restored percept, but it tends not to create percepts without a very strong base in the stimulation that is present before and after that stimulation. In the left-hand case, the required frequencies for reconstructing the full glide were actually present as part of the noise. In fact, you could argue that the glide itself was there, in the same sense that a statue is present in an uncut piece of stone. In the case of the auditory system, the process that acts as a sculptor, cutting the restored sound out of the available sensory stimulation, is the hypothesis that has been activated by the parts of the sound that occur on both sides of the interruption. The neural activity during the masker is not required to match the characteristics of the remaining parts of the interrupted sound themselves; rather it must match the properties of the missing parts of the hypothesized signal.

Further evidence that the perceptual restoration is required to be consistent with the neural stimulation in the B interval comes from experiments on phonemic restoration, in which a spoken utterance is interrupted by one or more bursts of louder sound. If the loud burst is fairly short, the listener does not perceive that any of the speech is missing; however, if the burst is removed and a silent gap is left in its place, the listener is aware of the gaps and their distorting effect on the speech.

It has been reported that the best restoration is obtained when the burst supplies the right kind of sound.[459] Richard Warren has argued that this happens "when the extraneous sound in a sentence is capable of masking the restored speech sound."[460] As I said before, I do not think that the auditory system needs to have an understanding of what would mask what. It would be sufficient for the system to simply know what sort of sensory evidence is needed to support the hypothesis that some particular sound, A, is present. This is a type of knowledge that it needs anyway if it is ever to recognize it. Then it would analyze the neural stimulation during the B interval for the sensory components signaling the presence of A. These components would always be present in B whenever B contained a masked A or even if B was composed only of a loud sound capable of masking A.

It turns out that the two explanations (masking potential versus hypothesis-confirming adequacy) are hard to distinguish empirically. The reason that the B sound would mask a speech sound is just because it activated those frequency-specific neural circuits that responded to that speech sound. Unless it did so, it would not mask the speech. Therefore, we know that the burst will create sensory activity that is characteristic of the missing sound (of course it will contain much more besides, as we argued with the metaphor of the uncut stone). As a result, it is hard to determine whether it is the masking potential of the signal or the fact that it includes hypothesis-confirming stimulation that enables the restoration to occur. I favor the latter hypothesis because it requires less knowledge on the part of the auditory system and because it is consistent with other examples of scene analysis.

If we assume that it is not the masking potential of B but its hypothesis-confirming content that is most important, the hearing of illusory continuity becomes just a particular case of hearing a continuing sound, A, when another sound enters the mixture. In all such cases we receive the sensory stimulation for A and then a mixture of the sensory effects of A and the other sound.

A simple case is one that we mentioned earlier: A band-limited noise burst, A, is alternated with a wider-band noise burst, B, that contains the frequency spectrum that is present in A (at the same intensity) as well as other frequencies, higher than those in A. If the two sounds are repeatedly alternated with no gaps between them, we will hear A continuing unchanged through B. In addition we will hear the higher frequency band of B, the part not present in A, as a sound that beats alongside the continuous A. In this example, we would not call the continuity of A an illusion. A is really present in B. The only reason for considering the demonstration interesting is the way in which we perceive the evidence coming from B. When B is played alone, it has the characteristic quality of a wide-band noise. However, we never hear that quality when A and B are rapidly alternated. Instead of hearing A–B–A–B–A–B– . . . , we decompose B, giving part of it to A and hearing the residual as a distinct new sound with its own high-pitched property. This result is really no different than what happens in any example of the continuity illusion in which a part of the stimulation received in the B interval is allocated to A. The only difference is that in most examples, we do not usually notice what happens perceptually to the rest of B when A is scooped out. This may be because the residual usually has no coherent structure of its own by which it can be recognized as different from the total B

sound, or because it is loud enough that "borrowing" a bit of sensory evidence with which to construct the missing part of A makes little difference in the global quality of B.

The fact that a pure tone can continue through a noise burst tells us something about what it takes to confirm the existence of A inside B. The frequency component of the noise that will act as the continuation of B need not be a resolvable component of B. Indeed, if it were independently audible (as in the case where, in listening to B alone, we could discern a tonal component at the frequency of A) we would not think of the continuity of A during B as an illusion. It is because the component is not separately audible when B is presented alone that we think of the phenomenon as an illusion. Apparently the existence of A before and after B allows us to "scoop out" some of the energy of B as a separate sound. This is not unlike the examples that we saw earlier in which a complex tone, B, was alternated with a pure tone, A, whose frequency matched one of the harmonics of B. The alternation allowed the listener to hear out the embedded harmonic. Yet we did not think of it as an illusion.

If we look at the intensity of B that is required before restoration of A will occur, we see that the amount of energy that is produced by B in a single critical band is quantitatively sufficient to support the hypothesis that A is really present. For example, suppose we have a 1,000-Hz tone that is alternating with a band-passed noise that has a flat spectrum within the band. The critical band at 1,000 Hz is about 175 Hz.[461] Suppose the noise bandwidth is 3,000 Hz. That means that about 5 percent of the energy of the noise is received in this critical band. Converting to decibels, the critical band receives 13 dB less energy than the total energy of the noise. This means that when the tone is 13 dB less intense than the noise, the critical band at 1,000 Hz will get the same amount of energy from the tone and the noise. The intensity difference between the tone and noise that is used in experiments with these signals is typically about this size or larger.

The requirement that tone B must supply enough energy in the critical band occupied by A is equivalent to the requirement for perceived continuity that was stated by Houtgast: "When a tone and a stimulus S are alternated (alternation cycle about 4 Hz), the tone is perceived as being continuous when the transition from . . . tone to S causes no (perceptible) decrease of nervous activity for any frequency region."[462] Houtgast's criterion would also guarantee, by the way, that the offsets and onsets in A would not be audible, since sufficient energy at the frequency of A would continue right through B; in effect, there really would be no offset or onset "edges" of A.

Experiments that were done on the pulsation threshold by Houtgast are relevant to these issues. When a pure tone, A, is alternated with a much louder complex tone, B, with no gap between them, the pure tone seems to continue right through the complex one. If A is gradually raised in intensity, a point is reached at which it no longer seems continuous but seems to go off or be reduced in intensity during B. This makes A seem to pulse periodically, and is therefore called the pulsation threshold.[463] This can be thought of as the point at which the auditory system can tell that there was a reduction in intensity at the frequency of A when B came on.

If B has a harmonic at or near the frequency of A, then A will sound continuous up to a much greater intensity (as much as 20 dB higher) than it does if there is no matching harmonic in B. This makes sense. Since there is more energy in B at that frequency, the ear will not detect a drop in energy at the juncture between A and B. Therefore it will not hear A as dropping in intensity. The intensity of A will have to be raised before this difference will be detected. For this reason, the pulsation threshold can be used to study how well the auditory system resolves harmonics: If there is no difference in the threshold as A deviates more and more from the frequency of a particular harmonic of B, this means that as far as the ear is concerned the spectral regions near the harmonic in B have as much energy as the region that is right at the harmonic. In other words, the ear shows no frequency resolution in that range of frequencies.

What is interesting about this in the present context is that even when B has no harmonic that matches A, there is some soft level of A at which it sounds continuous through B. Which energy, then, from B is being matched up with A to create a perception of a continuous tone? Rather than thinking of matching energy at a particular frequency, we should think of matching neural activity. If there is spread of excitation from one basilar membrane region to another (whether this spread is mechanical or neural), then the auditory system, when it detects a certain level of neural activity at a particular frequency region, will not know whether this occurred because of stimulation by a signal at that frequency or because of spread of excitation from a stronger signal at some other frequency. Therefore when we say that tone A has captured a component of B we are being imprecise. We should actually say that the neural activity stimulated by A has captured some of the neural activity stimulated by B. In natural listening situations, this pattern of neural activity (a part of the activity at time B matching the activity at time A) will usually exist because the event that stimulated the neural activity during time period A has actually continued during time period B.

We should remember, however, that the neural activity in B that is "captured" out of it need not match A exactly. This is supported by one of Helmholtz's observations. I mentioned it earlier when I cited his argument that the perception of the matching component in B was not an illusion. The observation was that even if the component inside B did not match A exactly, it could be captured by A, yet the listener would be able to hear that A and the extracted component were different. The utility of this spread of capturing for scene analysis is that even if a sound changes slightly in pitch over time, it will be able to be tracked into a mixture.

The "A1-A2 Grouping" Rule

Another rule that I have proposed as governing the illusion could be called the "A1-A2 grouping" rule. This says that A1 and A2 will be treated as parts of the same sound only if the rules for sequential integration cause them to be grouped into the same stream. This means that they would have been grouped as parts of the same stream even if there were no B sound between them. Auditory stream segregation and the continuity illusion have been described in the literature as separate phenomena. However, both can be understood as arising from the scene-analysis process. Stream segregation occurs when the auditory system attempts to group auditory components into streams, each stream representing a single external source of sound. The continuity illusion occurs when the system interprets a sequence of acoustic inputs as the result of a softer sound being interrupted by, and masked by, a louder one.

An argument can be made that the illusion of continuity must depend on the processes that produce auditory stream segregation. It goes as follows: The continuity illusion depends on the interpretation that one sound has interrupted another. This means that A1 and A2 have to be interpreted as parts of the same sound but not as part of the same sound as B. Unless A1-A2 hangs together as a stream, there is no sense to the statement that B has interrupted it. Continuity through interruption is a special form of stream integration that occurs when it is plausible to interpret A1 and A2 not only as part of the same stream, but as a single continuous event within that stream. Therefore the requirements for the continuity illusion *include* those for stream integration.

There are two consequences of the assumption that restored continuity depends on stream organization. Since stream cues are part of the "interruption" decision and not vice versa, cues for continuity (or discontinuity) should not make a difference to stream segregation, but stream cues should make a difference to continuity.

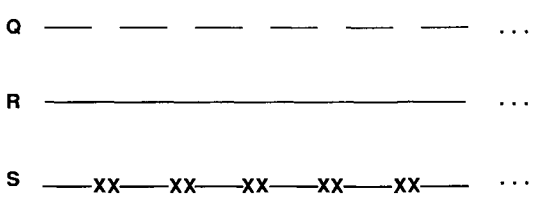


Figure 3.26
Streams, units, and continuity.

This means, for example, that adding a loud masking noise burst between two sounds, A1 and A2, should never affect the way that they enter into streams. If the stream-building process would have assigned them to different streams in the acoustic context in which they are situated, then the addition of a masking noise between them will make no difference and they will not be treated as two parts of a single sound. What, then, should the B noise do? According to this hypothesis, it should affect only those cases in which A1 and A2 would have been allocated to the same stream. But in those cases it should add the following element to the description: Not only have A1 and A2 been emitted by the same source but they are parts of a single sound from that source.

Figure 3.26 can clarify this. Let us distinguish three sounds, labeled Q, R, and S. Q is a tone that is pulsing on and off eight times per second, R is an unbroken 10-second-long tone, and S is the same as Q except that the silent gaps in the bursts of tone are replaced by loud noise bursts symbolized by the X's. In all cases, because all the parts of the tone are of the same frequency, the frequency-proximity rule for scene analysis will assign them all to the same stream. However, in Q the bursts will be heard as separate events within the same stream. The addition of the noise bursts in S makes us hear it like R, where all the segments of the tone are heard as one long event.

These sounds illustrate the fact that there are two aspects to the partitioning of a mixture of environmental sounds. One is to factor out of it the acoustic energy that arose from the same source. Another is to partition the energy of that source into separate events. An example of the first would be to group the sounds of a violin into a single stream, distinct from all co-occurring instrumental sounds. The second would involve hearing each note as a distinct event. Only by accomplishing both levels of grouping could we appreciate the rhythmic pattern of the notes. If the notes were strongly segregated into different streams, we would reject the rhythm as accidental. At the same time, unless there were distinct events starting at different

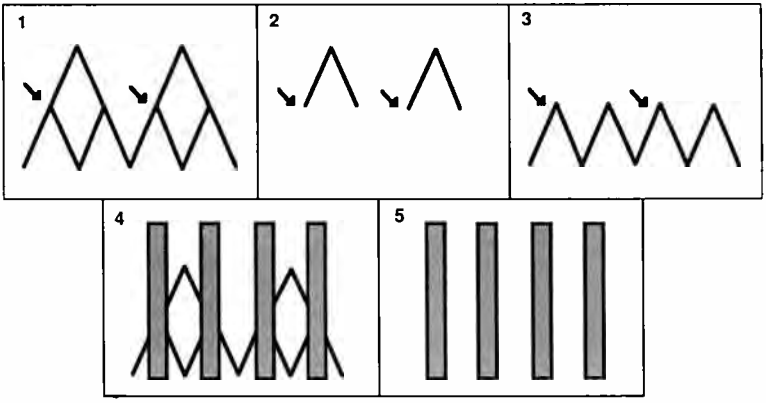


Figure 3.27
Pattern of glides used by Steiger (1980) to study the effects of perceptual restoration on stream organization.

times, there could be no rhythm in the stream. Music lends itself to convenient examples, but the situation is no different for everyday sounds. We could use a walker's footsteps in this example and draw the same conclusions.

In an informal experiment, Steiger created the acoustic pattern shown in part 1 of figure 3.27 in which a pattern of glides branches apart at what Steiger referred to as "decision nodes" (marked by arrows in the figure).[464] These patterns were segregated into two streams as shown in parts 2 and 3. Apparently when a stream reached the decision node it incorporated later material on the basis of frequency proximity rather than on the basis of "good continuation" of glide trajectories. Steiger asked the question as to whether the form taken by the stream segregation was related to whether the decision point was actually present or was, instead, created only by perceptual restoration. To answer it, he deleted the 20-msec portion just bracketing each decision point and replaced it with a loud white noise burst (part 4). The resulting streams sounded identical to those produced by the stimulus of part 1. There were three streams of sound. One consisted of the pattern of part 2, another consisted of the noise bursts shown in part 5, and the last contained the cycles of the lower stream, shown in part 3. Since this last stream did sound continuous, Steiger knew that perceptual restoration had occurred. However, it appeared that the restoration was not performed by extrapolating the glide from the the segment prior to the gap, but rather was based on information on both sides of the gap. The organization into two

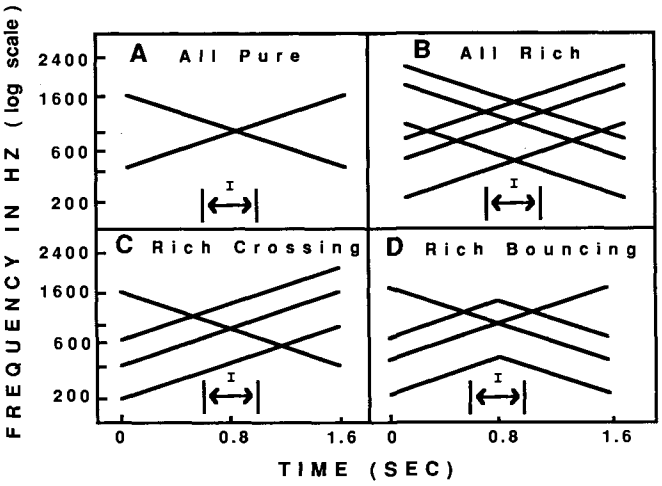


Figure 3.28
The four glide patterns (A–D) used by Tougas and Bregman (1985b). Interval I could be filled by glides (as shown), noise, or silence.

streams was the same whether or not the decision nodes were present or merely restored perceptually.

An experiment done by Yves Tougas and myself also showed that the perception of continuity through noise does not affect the organization into streams.[465] The auditory patterns that we used were the patterns of crossing glides that are shown in figure 3.28. Let us examine the perceptual results that occur when glides of this type are presented. In panel A we have an ascending and a descending glide. Each is a pure tone that is changing in frequency. Not surprisingly, this pattern will be heard as composed of two sounds. However, the two trajectories become ambiguous at the crossing point. Did the descending pure tone continue downward or turn around and go back up? One way that we hear this pattern is as a descending pure tone crossing an ascending pure tone. This is analogous to the way that our eyes tend to organize the diagram. It can be referred to as the "crossing" percept. The second way is to hear a higher glide falling to the midpoint and then "bouncing" upward again, accompanied by a lower glide that starts by gliding upward to the midpoint and falling downward again. This is analogous to visually dividing the figure into an upper V and a lower inverted V. It can be called the bouncing percept. The pattern in panel A tends to be heard in the bouncing mode because the principle of frequency proximity favors the grouping of glides that stay in the same frequency region.

In the pattern of panel C, the ascending glide is actually three related harmonics (the first, second, and third). We tend to hear this in the crossing mode because the auditory system fuses the ascending glide segments on the basis of their harmonic relations; therefore in the first half of the pattern it hears a pure descending glide and a rich ascending glide. Then at the midpoint, it seems to follow the same complex tone through the crossover point. The same seems to be done with the two parts of the descending pure glide. This grouping by harmonic complexity favors the crossing percept.

In pattern B, both the ascending and descending glides are formed of three harmonics throughout. Therefore the principle of grouping by the same harmonic pattern cannot favor either organization. The auditory system falls back on the tendency to group by frequency proximity just as it does in panel A. Again we hear the bouncing percept, this time involving two rich sounds.

In panel D, there is a very strong tendency to hear bouncing because the lower inverted V sound is enriched by added harmonics while the upper V is not. Therefore two tendencies favor the bouncing tendency—the tendency to restrict streams to a narrow frequency range and the tendency to group sounds with the same pattern of harmonics.

What does this have to do with the continuity illusion? There were two other sets of conditions based on the previous four patterns. One was made by deleting the middle 200-msec interval of each pattern (labeled I in the figure) and replacing it with a loud noise burst. These were called the "noise" conditions to distinguish them from the original "continuous" conditions. We were interested in the question of whether the perceptual grouping of these glides would be affected by what happened in the middle interval I. Another set was made by replacing the deleted material with a silence. These "silence" conditions were meant as a control for the fact that the noise condition differed in two ways from the continuous condition: by the absence of the gliding sounds at the crossover point and by the presence of noise. The silence condition differed from the continuous condition in only a single way.

The subjects were trained on the distinction between the crossing and bouncing interpretations of crossing glide patterns. Then they heard each of the experimental patterns and were asked to rate on a seven-point scale how easily they could hear either percept in that pattern. The major finding of the experiment was the striking similarity in the stream organizations regardless of the acoustic content of the interval I.

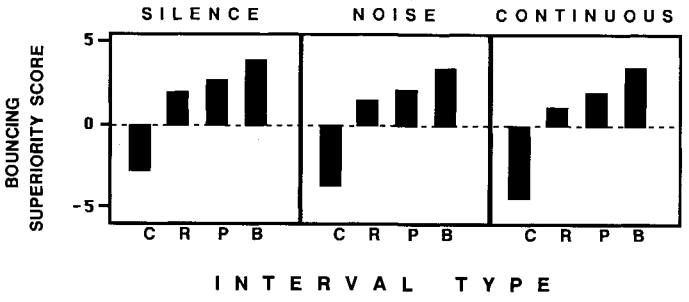


Figure 3.29
Data from Tougas and Bregman 1985b showing similarity of stream organizations despite illusory continuity.

Figure 3.29 shows the results, expressed in terms of a bouncing superiority score, which reflects the degree to which bouncing was easier to hear than crossing. Negative scores mean that crossing was the preferred percept. The three types of sound that occurred in the I interval (continuous glides, silence, and noise) yielded essentially the same stream formation and responded in the same way to the added harmonics. The only glide pattern in which there was a preference for the crossing percept (negative score) was the rich crossing condition (C) in which the added harmonics favored the crossing percept. The greatest preference for bouncing was shown in the rich bouncing condition (B) in which the added harmonics favored crossing. The conditions in which all glides were pure (P) or rich (R) showed a smaller preference for bouncing. The similarity of the results across conditions lends support to the idea that perceived continuity depends on the formation of streams and that the presence or absence of illusory continuity. The competition between grouping by harmonic pattern and grouping by frequency proximity resolved itself in the same way in all cases.

The results of this experiment appear to be explainable by a two-stage process. First, auditory streams were formed as a result of heuristics for grouping. Then, in response to cues for masking, a mechanism was activated to fill in the gaps in the already formed streams. According to this account, perceptual restoration works by the interpolation between events that lie within an already formed auditory stream and is a process that cannot affect the formation of such streams.

This experiment, by the way, and the informal demonstration by Steiger, are the only ones that have shown the perceptual restoration

of the continuity of two streams concurrently. Furthermore, they have done so in a situation in which the two streams themselves were created by stream-forming heuristics. Remember that the glides occurred simultaneously. Their segregation into two streams was done on the basis of differences in frequency movement of the ascending and descending glides.

To understand the significance of this demonstration, we should remember that in natural listening situations several sounds are usually occurring at the same time. When a loud sound occurs, it often masks two or more of the weaker ones. When the loud sound disappears again we would not like to find ourselves listening to a different signal than we were before the interruption had occurred. And we should not want to recognize chimeric sounds created by grouping the earlier part of one signal with the later part of another. This is why the stream organizing processes must not be changed by the processes that restore continuity.

To repeat an earlier point: My account of the relation between stream segregation and perceived continuity implies two consequences. The first, supported by Steiger's demonstration and by the experiment that Tougas and I did, is that interruption by a louder sound, leading to continuity, should not affect the perceived groupings. The second is that the rules for grouping *should* affect the restoration of perceptual continuity. Effects of this latter type have been found in research done by Valter Ciocca and myself.[466] The pattern that we used was a gliding pure tone (A1), then a loud noise burst (B), then another gliding pure tone (A2). There were no silent gaps between A1, B, and A2. The listeners were led to believe, by a series of pretests, that sometimes the glide continued behind the noise and sometimes did not. They were asked in the main experiment to judge whether the glide continued behind the noise. They used a nine-point rating scale that ran from "very sure not present" to "very sure present." In actuality, the gliding tone was never present with the noise.[467] We varied the frequency placement and the slope of A1 and A2. For example, A1 could be an ascending glide and A2 could be its mirror image, beginning at the highest frequency that glide A1 had reached and gliding back down to the initial frequency of A1. This condition led to a good restoration of the missing glide, presumably because the two glide segments were in the same frequency region and therefore their perceptual grouping would be favored by frequency proximity. In another condition, both A1 and A2 were ascending, with A2 starting at the exact frequency that A1 would have reached if it had continued with a constant slope behind B. This condition was

also judged as showing good continuity, presumably because the listener was able to project the trajectory of A1 through the noise. Other conditions, that were favored neither by frequency proximity nor by trajectory alignment, showed a weaker perception of continuity.

In this experiment, we found that two of the factors that are known to cause sequences of tones to be heard as a single stream also determined whether continuity would be heard. These factors are frequency proximity and alignment on a trajectory. Although the frequency proximity is always found to be effective in uniting sequences of sounds, alignment along a trajectory is not always found to do so.[468] In chapter 4, I have decided to consider the tracking of trajectories to be a process of schema-based stream segregation, involving attention to, and familiarity with, regular patterns of sound. Whatever the theoretical explanation for it, though, we know that in some cases, we can extract known or regular sequences out of mixtures more easily than unknown or irregular ones. Similarly, the restoration of the missing parts of sequences that display a regular pattern is widely found.

The presence of A2, appearing as soon as B has disappeared, seems to be of great importance to the continuity illusion; otherwise A tends not to be heard as continuing behind B. Giovanni Vicario has stressed the retrograde fashion in which the continuity illusion is controlled, with the mechanism deciding after the fact whether it heard the softer sound during the louder one.[469] This shows that the "whether" decision can be influenced by the sound that comes after the interruption.

The "what" decision is also influenced by material coming after the interruption. This has been demonstrated in phonemic restoration. When a speech sound is replaced by a loud noise, the sound that is perceptually restored can depend on material that comes after the noise.[470] For example, it has been reported that in the sentence, "It was found that the *eel was on the orange" (where the asterisk represents the sound of a loud cough that replaced a speech sound), the listeners would hear the incomplete word as "peel". If the final word was "table", then they would hear it as "meal". If the final word was "axle", the restored word would be "wheel".[471]

Obviously, from a logical point of view, the evidence in the B-A2 transition or in the material of A2 should be every bit as good as that during A1 or at the A1-B transition. However, we tend to believe that we hear a thing happening when it really is there rather than later when the evidence is all in. Our resistance to believing in retrospective effects partly arises from our belief that if some new element

entered our experience sometime after the event, we would have to push our perception of everything after this element backward or else distort the timing relations. Perhaps a metaphor will help resolve this problem. Imagine a musician trying to transcribe a simple melody into musical notation as a person sings it. He is always operating with some delay and sometimes the delay can get longer than at other times. Despite this, the "score times" written on the paper can always be correct, regardless of the particular moment in "transcription time" that they happen to have been written. If the musician were transcribing a two-part melody, the corresponding notes in the two parts could be in correct registration on the score even if they were not written at the same moment. If we consider that the musician in this example represents our active perceptual processes and the score represents the final product of perception, we can see how our perceptual decisions can be made at variable periods after an event has occurred, but our perceptual record of the events need contain no distortions.

The "A Is Not B" Rule

So far we have shown how, in deciding about the continuity of sounds, the auditory system gathers evidence that a sound A has been interrupted by a second sound B. We have just argued that this interpretation of the acoustic pattern will not occur unless A1 and A2 are perceived as parts of the same sound. There is a also second requirement: B must not be interpreted as part of the same sound as A1 and A2. Otherwise we would hear A1 turning into B and then B turning into A2.

Both the interpretations "A transforming into B" and "A interrupted by B" are logically possible when we hear the acoustic transition between A1 and B. How could we decide which has taken place? I would propose that the interpretation that the sound has undergone a transformation requires the change to be slow and continuous. An example of a transition that could be interpreted either way, depending on the abruptness of the transition, is the transition from a soft sound to a loud one with the same structure and back again. From our everyday experience we know that it is possible to hear a sound get louder and then soft again. In such a case we do not hear the softer version as a separate sound continuing behind the louder one. Yet in the laboratory, we can alternate two sounds whose only difference is one of loudness and hear the softer one continue behind the louder.[472] But in this case, there is an abrupt transition between the softer and louder versions and back again. We also know that in the laboratory we can induce illusory continuity by alternating a soft pure tone with

another pure tone that is louder and at a slightly different frequency.[473] Yet we know that if this transition were gradual, we would hear a single sound undergoing simultaneous frequency and amplitude changes.

We have observed a similar treatment of changes when we perceptually segregate tones that are alternating between different frequency regions into separate streams. When the high and low tones are joined by frequency glides instead of changing discretely in frequency, they are more likely to be incorporated into the same stream.[474] It is as if the auditory system does not believe that a sound can be transformed into another one instantaneously without going through intermediate stages.

If the preceding reasoning is correct, one of the crucial elements in the continuity illusion is that B should be segregated from A1 and A2 and put into a separate stream. This agrees with our experience of illusory continuity. The segregation of A and B into separate streams can be observed most easily in the version of the illusion in which a soft A and a louder B are repeatedly alternated; we hear B as a separate sound source whose bursts are superimposed on a continuous A.

It is likely that in the cases in which the continuity illusion occurs, a sudden change in intensity acts as if it were triggering two rules in the auditory system. Rule 1 is that the analysis of the current sound should be suspended and a new stream started. Rule 2 is that the new stream should be connected to some former stream if possible, and if the connection is successful, the analysis of that continuing stream should be resumed.

Rule 1 occurs at the boundaries between A1 and B and between B and A2. Rule 2 binds all the A's together and all the B's together. Since the changes in loudness of A at the boundaries is masked and some neural evidence for its continuation occurs during B, A is heard as continuous through B. However, since B is louder (and often spans a wider frequency range) its amplitude transitions are heard; therefore it is not perceived as a continuous sound but as a discontinuous series of bursts coming from a common source.

Let me make a brief digression. The idea that a sudden discontinuity induces the auditory system to begin a new analysis is supported by an informal observation that I made with the assistance of Pierre Abdel Ahad at McGill. We created a sequence of three tones with abrupt rises and gradual (exponential) decays. The decays were long enough that the notes overlapped one another substantially in time. When played in the forward direction, we heard a sequence of three tones with sharply defined pitches. However, when we played the sequence backward so that the notes had gradual onsets and abrupt

decays, the sense of three definite and distinct pitches was lost and the pitches blurred together.

As another brief digression I would like to propose an experiment on the abruptness of spectral changes. It should be possible to set up a pattern in which a pure tone A turned continuously into a wide-band noise B by spectral widening and then back to A again. The B-A2 transition could be created by passing a noise burst into a resonator, and then feeding the output of the resonator back into itself recursively until it became a pure tone. The A1-B transition could be created by recording the A-B transition and then playing it backward. Splicing the two together would create an A1-B-A2 sequence that, if the theory is correct, would not yield illusory continuity of A.

The "Old-Plus-New" Heuristic

Earlier I argued that the continuity illusion was analogous to following a sound into a mixture of sounds and factoring it out of the mixture. This is a topic that we discussed extensively under the heading of spectral grouping. We saw that when a mixture is decomposed by the heuristics of primitive scene analysis into a continuing sound and a residual, this affects our perception of the residual as well. It appears to have properties that are determined by only the acoustic components that are left behind after the parts that match the continuing sound are taken away. This is what we called the old-plus-new heuristic. If illusory continuity is merely another example of this process of decomposing mixtures, it too should show these effects on the perception of the residual.

There is just one qualification to be added to this prediction. Later I will argue that the creation of a residual with its own properties will occur only when the mixture has been decomposed by primitive, as opposed to schema-governed processes of scene analysis. The primitive processes are conceived to be innate and not based on learned properties of the signal. Therefore, I would imagine that when an interrupting sound B is interpreted as having a part that matches A, this will give rise to a residual, with properties different from the total sound, but only in certain cases. The residual will be created whenever the decomposition has been based on primitive features of the sound, such as the feature of occupying a particular frequency region.

We have already examined two cases in which a definite residual is formed. One occurs when a soft burst of noise is alternated with another one that is much louder but has the same spectrum. We hear the softer sound continuing through the interruption, and hear the residual of the loud sound. We can show that a residual has been formed in this case because the interrupting sound seems less loud

than it would be if it were presented alone; therefore some of its energy must have been allocated to the continuation of the softer sound. The second clear case of the formation of a residual occurs when a low-pitched narrow-band noise band is alternated with a wider band of noise that contains both the frequencies of the narrower band plus higher ones. We hear the residual as a series of high-pitched noise bursts accompanying the continuous lower-pitched noise. In both cases, not only is part of the weaker signal restored, but the stronger signal is heard differently. This shows that the restoration of continuity in the weaker signal is an example of the old-plus-new heuristic encountered in earlier examples of spectral decomposition.

In the cases that I have discussed so far, only the simplest characteristics of the weaker signal were used to extract it from the louder one. Yet there are other cases in which learned knowledge of the signal is involved in the extraction of the continuing signal from the noise. One example is the phenomenon of phonemic restoration, where a sentence is heard as continuing through an interrupting loud noise, despite the fact that the noise actually has replaced deleted phonetic material. Clearly in this instance the restored sound is not simply a copy of what went before the noise, but an extrapolation based on the listener's knowledge of what a speaker is likely to be saying. It is therefore based on schema-governed stream segregation and for this reason will not give rise to a distinct residual. There are other restorations clearly based on learned properties of signals. For example, it has been reported that musical scales can be restored when one note is removed and replaced by a loud noise burst.[475] We would not expect a separate residual to be formed in this case either.

Examples Examined in the Light of Theory

We have spent some time in arguing that the continuity illusion was the same process that is responsible for the decomposition of mixtures. It would be interesting now to use this approach to interpret the kinds of illusory continuity that have been observed. I will follow the order, from simple to complex, that is given in Richard Warren's presentation.[476] It appears that any sort of sound at all can be restored when a loud noise replaces a brief segment of it. We will see that in all cases, the neural stimulation that is characteristic of the restored sound can be found as part of the stimulation provided by the interrupting sound. This means that only a decomposition of this stimulation is required to restore the percept of the weaker sound.

The simplest case is what Warren refers to as "homophonic continuity."[477] This occurs when two sounds of different loudness

but identical spectral content are alternated. One example that has been studied is the alternation of 300-msec bursts of 70 and 80 dB intensities of a one-octave-band burst of noise centered at 2,000 Hz. The weaker sound seems to be present continuously, continuing through the louder. Another example is obtained when two tones of the same frequency but of different intensities are alternated.[478] Again the softer one seems to continue through the louder one and to be present all the time. In these cases, it is obvious that the evidence required to create the weaker sound is embedded in the sensory stimulation arising from the stronger one.

A slightly more complex case occurs when the soft sound (A) and the louder one (B) are different in spectrum as well as in loudness. The earliest report was in 1950 by Miller and Licklider, who alternated 50-msec bursts of a pure tone and a broad-band noise, and reported that the tone seemed to be on continuously.[479] They called it the "picket fence" effect, drawing an analogy with our visual experience of seeing a scene through a picket fence. In the visual case, the scene appears to be continuous and to complete itself behind the obstructing parts of the fence. In a similar way, the tone completes itself behind the interrupting noise bursts. This effect was independently discovered by Giovanni Vicario, in Padova, Italy, who called it the "acoustic tunnel effect" by analogy with the visual tunnel effect.[480] In this visual effect, a moving body passes behind a screen (the tunnel) and, after a time delay, a second one emerges from the other side. If the speed of the two motions, the delay between them, and the positions at which they enter and exit from behind the screen are appropriate, the viewer sees one single object in continuous motion pass behind the screen and come out the other side. The difference between the picket fence arrangement and the acoustic tunnel was that in the former the interruptions were periodic while in the latter there was only one. However, in both cases, restoration was found, and since then the requirement that there be sufficient neural stimulation in the appropriate frequency region has been shown to apply in both cases.[481] This requirement is consistent with the assumption that it is spectral decomposition that restores the continuity.

In the cases that I have just mentioned the restored signal is contained in the frequency spectrum of the louder one. This, however, is not always true. In 1957, Thurlow found that continuity could be obtained when alternating a weak pure tone (A) with a louder one (B) of a different frequency.[482] He called this an auditory figure-ground effect, by analogy with the Gestalt principle whereby visual displays are organized into figure and ground, with the ground appearing to be continuous behind the figure. When tones A and B are of different

frequencies, B does not contain the frequency content of tone A. However, if tone B is not of the same frequency as A its loudness must be increased to compensate for this difference in frequency. We know that the neural effects of a loud sound spread to adjacent frequency pathways in the auditory system and that this spread is greater with louder sounds. This leads to the conclusion that although B need not contain the frequencies of A, it must stimulate pathways that normally respond to A.

In the simpler forms of illusory continuity that we have just finished describing, the restoration can be thought of doing something fairly simple: continuing the perception of the sound that preceded the interruption in the exact form that it existed prior to the interruption. In the more complex form, however, the auditory experience that is restored must be thought of as a prediction from, rather than a continuation of, the interrupted sound. For example, we know that the missing parts of tonal glides can be restored. Gary Dannenbring, at McGill, created a pattern of alternately rising and falling glides.[483] We used this example in chapter 1. His stimulus pattern was shown in figure 1.15. The glides were interrupted by noise bursts either at the top and bottom vertex where the glides turned around, or else in the middle portion of each ascending and descending glide. Restoration was obtained in both cases.

We might imagine that this result suggests that the auditory system was extrapolating the glides into the noise. However, there is reason to doubt this conclusion. When the noise bursts replaced the center points of the ascending or descending parts of the pattern, the restoration was what you would expect, an illusory glide that joined the parts that were really there. However, when the noise burst was placed at the vertexes, replacing the turnaround point, the trajectory was not projected to the actual missing turnaround point. In one of Dannenbring's experiments, there was a sequence in which the noise replaced only the upper turnaround point. The listeners were presented with this as a test pattern that they were to listen to but were also given a second one that they could adjust. This adjustable pattern had a similar up-and-down gliding pattern, but no portions were missing, and the listeners could adjust how high the glide pattern swept in frequency (its lowest point was the same as that of the test pattern and its highest point had a brief steady-state portion of the same duration as the noise in the test pattern). They were asked to set the adjustable pattern so that it had the same highest frequency as the test pattern appeared to have.

The results showed that the highest frequency that the listeners heard in the test pattern was not the frequency that would have been

there if the missing portions of the glides had been restored but was, instead, close to the highest frequency that *remained after the peaks were deleted*. The listeners never heard glides that appeared to have frequencies that were outside the range of the glides with which they were presented. That is why I said earlier that the restoration was based on interpolation rather than extrapolation. By following the slopes of the glides on both sides of the interruption, the listeners could logically have extrapolated what the frequency would have been at the missing vertex. However, they did not do so. Instead they simply interpolated between the frequencies before and after the noise burst. Another indication that the listeners did not extrapolate the glide trajectory into the noise was evident when Dannenbring and I were listening to some pretests of this pattern. What we did was simply to play the glide pattern over and over again, cutting away more and more sound on both sides of the peak and replacing it with noise. As we did so, the high pitch at which the glide turned around seemed to get lower and lower. If we had been extrapolating on the basis of the audible portion of the glide, this should not have happened. There was enough glide left in all cases for a successful extrapolation to the "true" vertex. A machine could have done it. Human ears, however, prefer to hear the high point dropping as more and more of the high-frequency portion of the pattern is deleted.

This result does not necessarily imply that the auditory system is incapable of measuring the slopes of glides. I have already described some of the work of Valter Ciocca, at McGill, who also studied the illusory continuity of glides through noise and found evidence that the auditory system could compare the slope of a glide that emerged from noise with what it had been when it entered the noise.[484]

Continuity of Words and Musical Scales

So far the process that carved out the missing sound from the neural pattern during the interruption could be viewed as a fairly primitive one that used general "knowledge" about the world of sound. It is reasonable to believe that this general type of knowledge could be wired into the auditory system. The examples of musical scale restoration and phonemic restoration, though, show that restorations can be based on prior knowledge of the stimulus.

Musical scale restoration works best when only a single note from an ascending or descending musical scale is replaced by a loud noise.[485] From this, you might think that all the auditory system would have to do is to simply insert a note halfway in pitch between the two notes on either side of the missing one. However, the notes of the major diatonic scale are not equally spaced on any physical

scale; so this solution would not work. The pitch of the missing note must be supplied from memory.

The involvement of specific knowledge is clearest in the case of phonemic restoration in which the restored sound is a word that is appropriate to the context of the sentence. Phonemic restoration has been widely studied. The first study of perceptual restoration of sounds, done by Miller and Licklider, used lists of spoken words as well as tones. Their study compared speech that was interrupted by periodic silent gaps (about 10 times per second) to speech in which these gaps were filled by louder white noise. While the noise made the speech appear smoother and more continuous, it did not make the listener more accurate in deciding what the words had been.[486] However, other research has since found that filling the gaps by noise can actually improve the accuracy of recognition.[487] Why should this be? After all, the nervous system cannot supply information that is missing from the stimulus. It can only supply its best guess. Why could it not do that without the noise filling the gap? I have argued earlier that the noise eliminates false transitions from sound to silence and vice versa that are interfering with the recognition of the sounds.[488] In addition, because silences are interpreted as part of the speech itself and not as an added sound, the rhythmic introduction of silences is heard as a rhythm in the speech itself and disrupts the listener's perception of any natural rhythms that may have been in the original speech.[489]

The method of repeatedly interrupting a stream of speech has been used to show how the restorations depend on being able to develop a guess about the deleted sound from a consideration of the other nearby words. An experiment was done by Bashford and Warren in which three types of spoken material were used: lists of unrelated words, a magazine article being read aloud, and the same article with its words typed in the reverse order. Obviously listeners should be able to predict the missing sounds only when they could make sense of the sequence, that is, only with the second type of material. The listeners heard a continuous reading in which 50 percent of the speech was chopped out and either replaced by a silence or by a noise that was about 10 dB louder than the speech. The deleted portions were strictly alternated with the intact portions at a rate that was controlled by the listeners. At very rapid alternations, they could not tell whether any portions of the signal were missing. They were asked to slow the rate down to the fastest one at which they could tell that bits of the speech were missing. When the interruptions were filled with silence they could tell that there was missing sound even at rates of 9

or 10 interruptions per second, and the nature of the spoken material had little effect. However, it did have a strong effect when the gaps were filled with noise. With the newspaper material read in the normal order, they were unable to tell that material was missing until the rate of alternation was slowed down to about one interruption every 0.6 second. This was about twice as slow as the rate required for the other two conditions. Apparently the listeners were able to use the context to infer what the missing material was and once the acoustics told them that a restoration was justified, the inferred material was inserted into their percepts in a seamless way. I would guess that with unpredictable verbal material, only very weak hypotheses were generated and these did not give the listeners the same impression that everything was there. A later study from the same laboratory also showed that prediction of the speech sound was involved in perceived continuity. It was found that the maximum length that the noise burst may be before continuity is no longer heard depends on the rate of the speech, with slower speech tolerating longer interruptions. In slower speech, there is less verbal material to be predicted and restored.[490]

Although many of the studies have interrupted continuous speech repeatedly, others have introduced only a single interruption of a word or sentence by a loud sound. The early studies of this type were done by Warren and his colleagues at the University of Wisconsin in Milwaukee.[491] Generally the results have been the same as for periodically interrupted speech, but the method, focusing as it does on a single restoration, makes possible a more incisive analysis of the process. For example, one study that used this technique revealed the fact that while the listeners could hear both the speech material and the loud noise, they could not report accurately where in the sentence the interruption had occurred. Yet in the case where the gap was not filled with noise the listeners were able to accurately judge which speech sound had been deleted.[492]

The reader may recall that when rapid sequences of sounds are segregated into different streams, the listener is unable to accurately judge the temporal relations between elements of different streams. The perceptual loss of the position of the loud noise relative to the speech sounds indicates that the speech and the noise have been segregated into separate streams, a condition that I proposed, earlier in this chapter, as a prerequisite for stream segregation. On the other hand, when the silent gap remained the listener could correctly judge its position. Apparently the silence was not segregated from the sounds. This makes sense ecologically. Silence should not be treated as some-

thing that can interrupt a sound. There is no physical source in nature that can broadcast silences, superimposing them on top of other sounds so as to interrupt them.

The technique of looking closely at the effects of a single interruption has also revealed other effects. I have mentioned one of these already: Words that come after the interruption can affect how the listener restores the missing phoneme.[493] Another is that the restoration will be consistent with the specific nature of the neural stimulation provided by the interrupting sound (in the same way as in the nonspeech examples that we examined earlier).[494]

Integration as the Default Condition in the Auditory System

We can obtain evidence, from the illusion of continuity, that is pertinent to a question we raised earlier: Is the "default" condition in the auditory system to segregate concurrent auditory components or to integrate them? The evidence from the continuity illusion supports our earlier conclusion that integration is the default.

If we interrupt a tone by a noise burst and leave the noise on indefinitely, we may continue to hear the tone continue for some time but it will not last forever. Why is this important? In the continuity illusion, it is apparent that the segregation of the tone's frequency components from the noise can be attributed to a single factor—the presentation of the tone alone prior to the noise. There need be no properties of the spectrum of the noise itself that segregate those components. The fact that the perception of the tone gradually fades away suggests that the default condition is integration rather than segregation. When the segregating effects of the isolated presentation of the tone recede into the past, the default integration takes over.

We could have arrived at the same conclusion in a different way. If segregation were the default condition and if we listened to a simple long burst of white noise long enough, it would eventually fall apart into narrow-band frequency components. This would occur because the simultaneous onsets of all the frequency components at the onset of the noise, telling us to integrate all these components, would fade into the past and no longer be effective in opposing the default condition of segregation. We know that this result is never obtained.

Having integration as the default makes sense, since without specific evidence as to how to break down the spectrum into parts, an infinite number of subdivisions can be made, each equally plausible, with no principled way to choose among them. The conservative approach is to treat the spectrum as a unitary block except when there is specific evidence that this interpretation is wrong.

Duration of the Softer Tone

The continuity illusion, like most other organizational phenomena in audition, is not an all-or-nothing affair. Even under circumstances in which the interruption is too long to allow the softer one to go right through it, the softer sound may appear to travel a little way into the loud sound and to be heard again prior to the cessation of the interruption. This effect has been found when a 1-second sinusoidal tone was alternated with a 0.5-second burst of narrow-band noise.[495]

The length of time that a sound that is replaced by a louder one seems to continue on behind the louder one varies a great deal. Some sounds continue for quite a long time. There is an anecdote about a music teacher who gradually increased the intensity of a noise source that accompanied a recording of a musical performance and asked his class to signal when they could no longer hear the music.[496] Unbeknownst to them, when the noise became very loud, he switched off the music. Many members of the class reported that they heard the music long after it was switched off. Warren and his associates reported a laboratory observation of a case of long-lasting continuity. The softer sound was a $\frac{1}{3}$-octave band of noise centered on 1,000 Hz and the louder one was a band of pink noise extending from 500 to 2,000 Hz. They always had the same duration (D) and were alternated. However, in different conditions, D was varied so that each sound could stay on for quite a long time. All their 15 listeners heard the narrower-band noise continuing through when the loud noise lasted for several seconds, and six of them could hear it for 50 seconds.[497]

In most cases of illusory continuity (for example, with pure tones continuing through noise) the continuity lasts much less time than this, often less than a second. Perhaps the difference depends on how hard it is to imagine A inside B. Remember, first, that a long noise burst is not a sound with unvarying qualities. It is random, and we can hear different qualities of hisses, rumbles, pops, and so on, within it at different moments as we continue to listen. If A is a clear and definite sound like a pure tone, it may be hard to imagine it in the continuously varying noise. However, if A itself is a sound with constantly varying properties, we may be able to pick out some of these often enough, within a long B noise, to sustain the imagined presence of A. On this hypothesis, it should be wide-band irregular sounds that seem to continue the longest, since these offer the greatest array of different properties to search for in the louder noise. This account is consistent with the idea that the search for A inside B is not a simple process that takes place at peripheral levels of the auditory system but at higher sites in the central nervous system. Of course we already

know that the higher sites are involved because they must surely be involved in phonemic restoration.

The "Roll" Effect

Another perceptual illusion related to illusory continuity was discovered by Leo van Noorden.[498] It also involves the alternation of a softer and a louder sound but is different from the continuity illusion in one basic way. There is a silent gap between the softer and louder sounds. Therefore, instead of the softer appearing to be continuous behind the louder, it appears first to go off (when it actually does), and then to come on again when the loud sound does. The effect has been found only with a very rapid alternation of two short sounds, the rate exceeding about 12.5 tones per second. Van Noorden used pure tones, A and B, that either were at the same frequency or at very close frequencies. To get the effect, you also have to make the two tones different in loudness and focus your attention on the softer tone, A.

Let us suppose that each tone is 40 msec in duration and there is a 10-msec silence between them. If the two are not very different in loudness, you hear a string of tones occurring 20 times per second. If A and B are very different in loudness, tone A disappears (through masking) and you hear only B occurring at 10 tones per second. The interesting effects occur at intermediate differences in loudness. Suppose, while alternating A and B, you start increasing the intensity of A from the point at which it is inaudible. When A first becomes audible, you will hear the continuity illusion. That is, a soft tone will be heard as being continuously present behind the 10 pulses per second of tone B. Then, as the intensity of A is raised further, you will begin to be able to hear pulsing at two different rates. If you turn your attention to the louder tone, you can hear it pulsing at 10 tones per second as before. However, if you turn your attention to the softer tone, it appears to be pulsing at twice this rate. It is as if the louder tone had split into two parts, one part contributing to the 10-Hz pulsation that you experience as a louder stream, and the other grouping with the 10-Hz pulsing of the softer tones to create a softer sequence with a 20-Hz pulse rate. This 20-Hz pulsation of the softer tone sounds like a drum roll; hence the name "roll effect."

There is another feature of the effect that is important. If you start with a condition that gives the roll effect and start to slow it down, eventually (say at rates slower than 12.5 tones per second), the extra pulses of the softer sound, A, will disappear and you will hear it pulsing at 10 Hz, the same as B. Furthermore, you can pick out either

the soft tone stream or the loud tone stream by listening for one or the other.

A third basic fact about the roll effect involves the frequency separation between tones A and B. If adequate conditions for obtaining the roll effect are set up (say at an overall rate of 20 tones per second), and you begin to increase the frequency separation between A and B, the roll effect will start to disappear and both the A and the B tones will start to be heard as pulsing at their own 10-Hz rates.

The roll effect seems to be a version of the continuity effect that differs from it in that the softer tone, rather than sounding continuous, seems to be pulsing, the onsets of the pulse being determined by both the onsets of the softer tone and those of the louder one. Because part of this pulsing is derived from the onsets and offsets of the weak tone itself, this implies that its onsets and offsets are not being masked. This idea is consistent with the observation that we can convert the roll effect into the continuity illusion by merely decreasing the intensity of the weaker tone. Instead of hearing the weaker tone pulsing at a high rate, we begin to hear it as continuous. Evidently the reduction of the intensity of the softer tone allows the masking of its onsets and offsets by the louder, a condition that, as we saw earlier, is always required before illusory continuity can occur.

The roll effect seems to be a special case of the illusion of continuity in which the onsets and offsets are not completely masked but, instead, add the sensation of extra onsets and offsets. This idea of a close relation between the continuity and roll effects is supported by the observation that the two effects respond in the same way to the effects of the frequency difference between A and B and to the speed of the sequence.[499] In both effects, as the two frequencies are moved apart, the effect disappears in favor of a percept in which the A's and B's are perceived as separate sequences of distinct tones. In both cases, as well, when the frequency separation has destroyed the illusion, it can be restored if you speed up the sequence by shortening any silent gaps that may exist between the A's and B's. The effects of frequency separation have already been explained for the continuity illusion. As A moves further away from B in frequency, two things happen as a result of the decreasing overlap in the neural effects of A and B: The discontinuities of A are less masked and there is less evidence for the existence of A underneath B. Similarly, in the roll illusion, the frequency separation makes the onsets and offsets of the two tones more distinct from one another, and also decreases the evidence for A in the neural stimulation derived from B.

The effects of speed in overcoming frequency separation seem to be due to an increase in the masking of the discontinuities in the weaker

tone. In the continuity illusion, we argued that silent gaps will allow a separate registration of the onsets and offsets of the two tones. Eliminating them makes the discontinuities less audible. In the roll effect, the discontinuities in the weaker tone apparently must not be too distinct either. When they are made so, either by shortening the silent gaps or by separating the frequencies, the effect breaks down and two streams of tones, pulsing at the same rate, are heard.

Under conditions in which we are completely unable to hear the discontinuities in the weaker tone, we hear it as continuous. If we hear the discontinuities as strongly segregated from those of the stronger tone, we simply hear two concurrent streams, soft and loud, pulsing at the same rate. There must be some intermediate status of our perception of the discontinuities, heard only under a very limited range of conditions, that yields the roll effect.

To summarize, the roll illusion, like the continuity effect, involves a perceptual synthesis of a second sound accompanying B. In both cases, the neural evidence required for the perceptually generated sound is actually present during the louder sound. In the roll effect, an onset of the weaker sound, accompanying the onset of the louder sound, is also added to the percept. In some sense the evidence for this onset is stolen from the evidence for the onset of the louder sound.

Comparison with Vision The role of the continuity illusion as a process of scene analysis can be clarified by examining both its similarities to, and differences with, some analogous phenomena in vision. The obvious analogy is a case in which one object obscures our view of a second one that is further away. Despite the fact that the visible parts of the further object are separated visually by parts of the nearer object they will be treated as continuous if their contours line up adequately and if they are similar enough, move in synchrony, and so on. If an object is moving, it can even be totally obscured as it passes behind another one and be experienced as continuously present to vision as long as the movement is continuous. This is the visual tunnel effect that we discussed earlier.

The similarities to audition are evident. The interrupter, whether it be object or sound, does not have the effect of dividing the interrupted thing into separate perceptual parts as long as the interrupter is seen as something distinct from the interrupted thing and if the separated parts of the interrupted thing fit together appropriately.

However, there are differences as well as similarities. They derive from the fact that most visual objects are opaque, but most auditory signals are transparent. Sounds are transparent because when two of

them are active at the same time, their properties are superimposed. In vision, in the case of opaque objects, if a nearer object covers another, only the properties of the nearer one are present to the eye in visual regions in which the nearer one covers the more distant one. These differences in the effects of "covering up" lead to different requirements, in vision and audition, for perceiving the continuation of one thing behind another.

Let us designate the interrupted sound or visual surface as A, and consider it to be divided into A1 and A2 by B, the interrupting entity. We can describe the differences between vision and audition as follows. In audition, B must be louder than A, but in vision B must be closer than A. In vision there must be a direct abutment of the surfaces of the screening and screened objects. One object's surface must end exactly where the other begins and the contours of A must reach dead ends where they visually meet the outline of B. In the auditory modality, the evidence for the continuity occurs in the properties of B itself as well as in A1 and A2; B must give rise to a set of neural properties that contains those of the missing part of A. In vision, on the other hand, if the objects are opaque, there is no hint of the properties of A in the visual region occupied by B.

In the case of partially transparent screens in vision, in addition to the "bounding edge" requirement there is the same requirement as in the auditory case for continuing evidence for the screened object. However, in the visual case, the continuing evidence for the screened object tells the viewer not only about its own continued existence but about the degree of transparency of the screen. In sound, transparency is not a property on which sounds vary (they are all equally transparent); so the continuing evidence for the screened sound does not specify any property of the screening sound.

Contralateral Induction

Earlier in this chapter, two phenomena were introduced as a pair—illusory continuity and contralateral induction. They were grouped together because they both involved taking neural activity that was stimulated by one auditory event and reallocating it so that it helped to define a second one. In examining the continuity illusion, we saw that some of the neural evidence provided by the louder sound was interpreted as belonging to the softer event.

Contralateral induction has been briefly introduced earlier. It involves the following: If we play a sound (let us call it the target sound) to one ear, it will be heard on that side of the listener's body. However, if we play a loud sound, different in quality from the first,

to the other ear at the same time (we can call this the inducing sound), the target sound will be heard as closer to the midline of the listener's body. The inducing sound will have induced an erroneous perception of the location of the target sound. In the example that I gave earlier, the target sound was a pure tone and the inducing sound was a noise burst. The explanation that I will offer, after presenting the evidence, is that if the neural activity stimulated by the inducing sound contains components that match those activated by the target sound, they will be allocated to the target sound. Since the target sound will now be defined by sounds arriving at both ears, it will be perceived as nearer to the center of the body than it was when unaccompanied by the inducer.

Why should we think of contralateral induction as a phenomenon worth mentioning? Aren't all cases of hearing with the two ears examples of contralateral induction? For example, if we played a pure tone of 1,000 Hz to the left ear of a listener, it would be heard on the left. Now if we played to the opposite ear, as the inducing tone, a pure tone of the same frequency, amplitude, and phase as the first, the target sound would now be pulled over to the midline of the listener's body. This simply follows from what we know about the interaural comparisons that take place in the normal localization of sound. Our special interest in contralateral induction seems to derive from the fact that the inducing sound is not the same as the target. If that is the case, how about the following example? We present a 500-Hz tone to the left ear and it is heard on the left. Then we play a mixture of a 500-Hz tone and an 1,100-Hz tone to the right ear. Depending on the intensity and phase of the 500-Hz component in the inducing tone, the 500-Hz sound will be pulled to the right (that is, toward the center of the body). At the same, the 1,100-Hz component will be heard as an isolated tone on the right. Again we are not surprised, because it follows from our belief that the auditory system can compare the neural activity at the two ears and find matches for individual frequency components. It need not treat the whole sound at each ear as a unit.

The novelty in contralateral induction is that there is no *specific* component or components at the second ear that have been inserted by the experimenter to match the target. Instead the match is due to some partial equivalence in neural activity at the two ears resulting from two sounds that we do not normally consider to have matching components. The extraction of some of the neural evidence activated by the inducing tone so as to match the target tone involves a decomposition of that evidence that would not have taken place if the inducer had not occurred at the same time as the target. For this reason,

while we cannot really think of this matching as an illusion, we can be interested in it as a scene-analysis process.

Early reports of this effect used stimuli in which speech was pulled to the center by noise played to the second ear or in which a tone was pulled by another tone of a different frequency or by a noise.[500] Similar induction effects have been been created by presenting a tone as the target sound over a single headphone and presenting an inducing tone or noise over loudspeakers.[501] In 1976, Warren and Bashford named the phenomenon "contralateral induction" and developed a method for measuring it.[502] Their innovation was to repeatedly swap the positions of the target and inducing sounds. For a half second, the target sound would be in the left ear and the inducing sound in the right; then for the next half second, the positions of the two would be reversed. In this manner, every half second, the positions would switch. It would be expected that the images of both the target sound and the inducing sound should oscillate back and forth between the two sides of the body since the signals on which they were based did so. However, this occurred only under circumstances where the induction failed. If the induction succeeded, only the inducing tone would seem to move. The target tone would seem to be on continuously and to stay near the center of the listener's body, or perhaps move only a little way to either side of the midline.

The experiment allowed the listeners to adjust the intensity of the target tone while the sequence of switches continued. The amount of contralateral induction depended on the relative intensity of the two sounds. If the target sound was substantially softer than the inducing sound, the induction occurred and the target sound seemed to remain in a somewhat vague position near the middle of the listener's body. However, when it was increased in loudness, a threshold point was reached at which the induction failed and the target sound seemed to oscillate from side to side, just as it would have if the inducing sound had not been present. In short, the induction required the intensity of the target to be low as compared with the inducer. The listeners were asked to adjust the intensity of the target to the maximum level at which it still did not appear to switch from side to side.

This method was used with targets that were pure tones ranging from 200 to 8,000 Hz and inducers that were noise bands filtered in different ways. The results are shown in figure 3.30. The experiment found that the induction seemed to work via the frequency components in the noise that matched the tone. For example, when the frequencies near that of the tone were filtered out of the noise (band reject condition), the intensity of the tone had to be turned down in order for induction to occur. On the other hand, if the frequencies

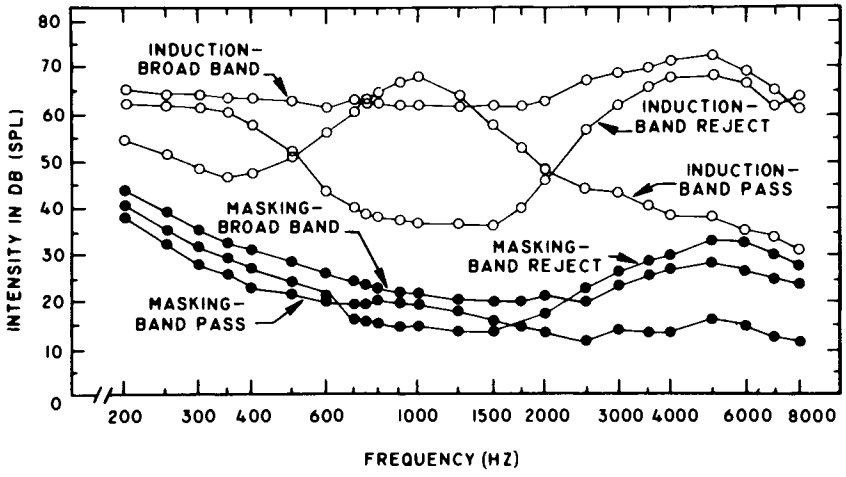


Figure 3.30
Upper intensity limit for contralateral induction of monaural tones presented with one of three contralateral 80-dB noises filtered as labeled on the graph (open circles). Detection thresholds in the presence of each kind of noise are shown as well (solid circles) (Warren and Bashford, 1976).

near that of the tone were retained but more distant ones were filtered out (induction band pass condition), the target could be turned up to a high level and still not overcome induction.

These effects resemble the continuity illusion in a number of ways. Let us adopt a terminology that helps to make this point. Earlier I referred to the tone that is heard as continuous in the continuity illusion as A but I used the word target when referring to the continuous tone in Warren and Bashford's version of contralateral induction. Let us call it A as well. Similarly, let us use the term B for both the loud interrupting sound in the continuity illusion and the loud inducing sound in contralateral induction.

The first similarity is that B must be much louder than A in both cases. Second, B must stimulate neural activity that is normally associated with A. Let us see what these two facts imply. In both cases, the auditory system is picking out from the loud sound, B, evidence that A is mixed in with B, and is finding this evidence because the neural response to A is a subset of the response to B. Therefore, both effects can be thought of as an attempt to deal with mixtures. Earlier the argument was made that one effect of the relative loudness of B was to mask the onsets and offsets of A. In Warren and Bashford's version of contralateral induction, there may be no

need for this. There are two phases to the cycle that they used, one in which A is on the left and B on the right and one in which their positions are reversed. If the neural activity that is required to specify the existence of A (the softer sound) is received in both ears in both phases, then there are no onsets and offsets in the A evidence in either ear. Therefore the masking of discontinuities is not relevant in this case. This is just as well, because the masking of a sound by one in the opposite ear is not very effective.

There is a second similarity. The continuity illusion is actually playing a role in Warren and Bashford's contralateral induction experiment. Consider what is happening at one ear. It receives a sequence of sounds, A–B–A–B–. . . , that is exactly the stimulus for the continuity illusion: an alternation of a weak sound with a stronger sound that evokes a neural response that includes the response to the weaker as a subset. One would imagine, then, that there are two influences that are helping to segregate the A activity that is embedded in the B activity: the fact that A precedes and follows B as in the continuity illusion, and the fact that A is occurring simultaneously in the opposite ear. The twofold influence on the segregation of the A component of B should stabilize the perception of A as centrally located.

One observation about the perception of A in this experiment was that its perceived location, although generally at the center of the body, was not experienced as precisely located in space. Why should this be true if the neural activity derived from A was being matched by activity derived from B? Should this not have given a definite location? Remember that B was a noise burst. The particular frequency components that were present and the phase of each varied randomly from moment to moment. Therefore, although in the long term, the spectrum of B had the components that were needed to match A, these components existed with random phases and amplitudes. Since the relative amplitudes and phases of corresponding frequency components at the two ears are what tell the listener the location of a sound, this irregularity of interaural relations would specify not a single location but a distribution of locations, having a central position that is given by the average of all the matches. This argument, however, would seem to imply a much wider dispersion in the location of A than we actually experience in contralateral induction. It is possible that the dispersion is limited by a tendency of the auditory system to average across rapid changes in the estimates of the spatial location of a sound. We will see evidence for such sequential averaging in an experiment by Steiger and Bregman.

A pure tone of one frequency can serve as the inducing sound when another pure tone, of a different frequency but much softer, is used as a target. This follows from the fact that the neural effects of a loud pure tone will spread out to adjacent neural pathways that are responsible for the detection of other frequencies. Therefore the auditory system will be able to detect sites of neural activity stimulated by the inducer that match sites on the target side. Of course, other sites on the contralateral side will also be active, but these will be rejected as the effects of a different sound.

In one of their conditions, Warren and Bashford used, as the target (A), a burst of noise and, as the inducer (B), a noise that was 1 dB louder than the target and had the same spectral characteristics. Being an independent sample of noise, however, it was uncorrelated with the target. The listener experienced this condition as containing two sounds, a continuous noise diffusely localized around the medial plane of the body and a second very faint noise oscillating back and forth and always localized at the ear that was receiving the louder sound. This effect reminds us of the homophonic continuity effect that we discussed under the topic of illusory continuity. We should, therefore, call the present effect "homophonic contralateral induction." In both effects the operation of the old-plus-new heuristic is displayed quite clearly. B is treated as a mixture of A and another sound, X, and X is heard as a separate sound that has the energy that would be required if X and A were to add up to give the total energy present in B. In the present case, since B is only 1 dB louder than A, the residual, X, has very little energy and is heard as a very faint sound.

The decomposition of the inducing sound (B) to derive a part that matches A is not just determined by the fact that A and B are present at the same time. This is fortunate; although there are times, in a natural listening environment, when the segregation should be done, there are others when it should not. It should be done in the following case: A is a sound emanating from a source that is straight ahead of us and therefore provides inputs to both ears whereas B comes from a source very near our right ear, so that the B energy mainly comes to our right ear. Let us further suppose that B has a spectrum that includes that of A, so that A is not distinguishable in the right-ear signal taken alone. In this case, since A and B are actually two environmental sounds, the A stimulation should be extracted from the B stimulation. Otherwise we would think that the A energy is present only at our left ear and that it comes, therefore, from a source close to that ear.

In the next example, the A sound should *not* be extracted from the B sound: A really does come from a source near our left ear and B

comes from a source near our right, and although there are frequency components in the right ear input that are the correct frequencies to have come from A, their source is actually not A. It just happens that B contains frequencies that match A. How can the auditory system tell which of these two real-world situations it is facing? The solution is to use the history of the sounds to decide.

As we have seen earlier in our discussion of the effects of stream organization on the localization of sounds, the perceived location of a sound does not just depend on the classical binaural cues. Continuity over time also plays a role in localization. Here is how we might use it in the present example. If A occurred alone close to our left ear and then, while it was continuing, B occurred at our right ear, a good strategy would be to continue to hear A at our left, despite the match of some of B's neural activity to A's. In other words we should use sequential grouping rules to group the energy arriving at the A ear so as to form a stream with what came earlier at that location rather than with what is happening at the B ear. In this way sequential grouping could correct the cases where binaural relations were deceptive.

Howard Steiger and I did an experiment to show that listeners could indeed make use of this sequential grouping strategy to prevent inappropriate contralateral induction.[503] The stimulus pattern included (in one time "frame") a pair of sounds, presented dichotically (one to each ear) over headphones as in the typical experiment on contralateral induction. The pair consisted of a pure tone at 1,024 Hz (the target) sent to the left headphone and a more intense one-octave noise band (the inducer), centered at 1,024 Hz, sent to the right headphone. The difference between this experiment and the typical one was that this frame alternated with a second frame in which there was only a single monaural sound, presented on the left. This monaural tone was called the captor tone because we expected it to capture the target tone out from within the dichotic tone so that it was no longer delateralized by the inducing noise. In other words, this sound was expected to tell the auditory system not to extract A from B's neural activity because A was merely a continuation of an earlier monaural event. The captor was always a pure tone which could vary in frequency.

The monaural and dichotic sounds were presented in a repeating sequence: the monaural sound for 100 msec, silence for 30 msec, dichotic sound for 100 msec, and then silence for 232 msec, the whole sequence repeated over and over. The listeners were asked to judge the left-right position of the target tone (and of the captor tone, in conditions in which it was present) by writing down a number

from 1, representing the leftmost extreme, to 12 representing the rightmost.

In some conditions the captor tone was omitted from the cycle. When this happened, the listeners judged the target as being near the midline of their bodies; the induction occurred normally. However, when the left-positioned captor tone preceded the dichotic tone, the perceived position of the target depended on the frequency of the target. When the frequency of the captor matched that of the target it partially prevented the latter from being pulled over to the middle by contralateral induction. It was as if the auditory system had decided that the target was a reoccurrence of the captor and therefore preferred that its location not be too different in the two occurrences. However, as the frequency of the captor was made to differ more and more from the target's, it had less and less of an effect in preventing induction. We should not be surprised at this fact. We already have seen, in our discussions of sequential stream segregation, that the strength of sequential grouping of two tones depends directly on how close their frequencies are.

We have seen, then, how sequential stream segregation competes with the across-ear relations to determine how the sound received at each ear is to contribute to localization. Later we will see that the competition can occur even when the contralateral sound is exactly identical to the target. In some cases, the competition is so extreme that the result is duplex perception.[504]

It is easy to set up simple demonstrations in which the scene-analysis process treats the signals at the two ears as defining separate sounds rather than allowing one to induce displacement of the other. In one example, we begin with a case of contralateral induction. We play a continuous soft tone of 1,000 Hz at the left ear and a continuous white noise at the other. As we slowly raise the intensity of the noise, we hear the tone appear to move toward the center of the body. We stop raising the intensity when the tone is fairly well centered (and before we do damage to our right ear). Now without changing the intensity of the noise, we change it from a continuous to a pulsing sound (switching abruptly between 0.3 second on and 0.7 second off on each cycle). Now the tone remains on the left of the body. The scene-analysis system has decided, despite the fact that it is receiving bilaterally matched 1,000-Hz stimulation when the noise is on, that the left-ear part of it really belongs to a continuing tone at the left. In other words, it has used the old-plus-new heuristic. Part of the evidence it has used is that the onsets and offsets on the right are not matched by any changes at the left.

A second example is even more dramatic. Instead of using a white noise as the inducing tone at the right ear, we use a tone of 1,000 Hz, exactly matching the one at the left in phase. When it is continuous, as we slowly raise it in intensity from zero, we hear a 1,000-Hz tone move toward the center of the body (and even past it if the right-ear tone becomes louder that the one on the left). Now leaving the right-ear tone equally as loud as the left one, we set it to pulsing just as we did with the noise. (However, the proportion of "on" time to the whole pulse cycle must be less in this example than for the noise of the previous example.) Again the scene-analysis system decides that the balanced and phase-matched stimulation received, when the right-ear tone is on, really belongs to two sounds, one at the right and one at the left.

The astute reader will have noticed that we have already encountered an effect where sequential stream segregation and binaural relations were set into competition. When we examined the effects of spatial location on sequential grouping, we discussed an experiment, carried out by Deutsch, in which listeners were asked to identify a short melody whose notes alternated between the ears. This was difficult, presumably because of stream segregation by location. The task became easier when a constant drone tone was presented to the ear that was contralateral to the ear that was getting the current tone from the melody.[505] I interpreted this as occurring because the equalization of intensity at the two ears reduced the cues for localization of the melody's tones and accordingly reduced their tendency to form separate streams. Yet Deutsch's drone was not at the same frequency as the tone in the melody. We have interpreted Warren and Bashford's research as showing that the delateralization occurs only to the extent that the contralateral sound stimulates neural pathways that correspond in frequency to those activated by the target sound. However, if we reexamine their data, shown in figure 3.30, we see that with the filtered-noise inducers, even in frequency regions that lacked frequencies that matched the target tone, contralateral induction was obtained, albeit at low target intensities. It seems possible that although the exact matching of frequencies plays a role in allowing part of the contralateral neural activity to be used to specify the location of a target, it is not essential. This lack of frequency specificity as exemplified by Deutsch's drone experiment may be due to a spread of neural activation across frequency regions at some low level of the auditory system or to some effects of contralateral energy that are not frequency specific and occur higher up in the system.

Perhaps we can summarize the contralateral induction effect as a laboratory phenomenon that, like the illusion of continuity, shows us

how powerful the auditory system is in finding frequency components that it is looking for even when this requires it to extract components out of mixtures. In the continuity illusion it is looking for frequency information that matches a previous signal, and in contralateral induction it is looking for a match to information received in the other ear. In both cases, it is able to carve what it needs out of the mixture, leaving the residual behind to be heard as a separate sound.

Our account of the phenomena of illusory continuity and contralateral induction has an important relation to the phenomenon of masking. Richard Warren has proposed that restoration will occur under conditions in which the interrupting or contralateral noise is loud enough to have masked the expected signal.[506] It has also been argued by many researchers that illusory continuity depends on the existence, during the interruption, of sufficient stimulation at the frequency of the expected sound to serve as evidence that the sound is really there. However, the two accounts are hard to distinguish because if a masker is to mask a target, then, in general, the masker must have about as much energy in the critical band (or bands) occupied by the target as the target itself has. An important point to remember is that in the stimulus for these two effects the evidence for the missing signal is not missing. It is simply part of a larger mixture of sound.

It is interesting to speculate about how this point relates to all other cases of masking. Auditory theory has usually considered masking to be a situation in which the masker has obliterated all neural evidence for the target. However, this may not be the best way to look at it. The better way may be to say that the evidence for the existence for the target has not been wiped out, but merely hidden. Let me offer an analogy. If you wanted to hide a red spot on a white canvas, the best way to do it would be to paint the rest of the canvas red too. The red spot would not be gone but the outline that formerly defined it against the white canvas would no longer exist. You can think of a masker as something that fills in the background in such a way that there is no longer any spectral shape defined against the white canvas of silence. The signal is still there, but it is camouflaged. Masking, then, is the loss of individuality of the neural consequences of the target sound, because there is no way to segregate it from the effects of the louder tone. If you introduce a strong basis for segregation, the signal can be heard as present. There are many ways that this could be done: You could expose the auditory system to earlier and later parts of the hidden sound as in the continuity illusion. You could expose it to sound in the other ear that matched part of the spectrum, as in

contralateral induction. You could expose the auditory system to a prior copy of the hidden sound as in the experiment by Bregman and Pinker.[507] Or, you could give the hidden sound a different location than the masker, or a different pitch, or different amplitude or frequency modulation. In other words, you could employ any of the factors that increase spectral segregation. It is for this reason that (as Rudolf Rasch has suggested about musical performances) masking in natural listening situations is less of a hindrance than laboratory studies would suggest.[508]

Summary

This chapter began with a demonstration that the neural evidence that results from an incoming spectrum is inherently ambiguous with respect to environmental sounds. Does the evidence tell us that there was only one sound with a complex spectrum or a number of sounds with simpler spectra? The rest of the chapter described a number of heuristics that the auditory system uses to decide how the evidence should be decomposed.

One of the main rules that the system uses is that if the neural activity evoked by an earlier sound resembles a subset of the current neural activity, that subset should be interpreted as due to the continuation of the earlier sound. Then the difference between the subset and the whole neural activity should be treated as a residual-evidence pool. This is called the "old-plus-new heuristic." The residual may be heard as a sound in its own right or be further broken down. This heuristic was seen as responsible for many well-known examples from the research literature, including the ability to hear out a component of a complex tone, or to hear a softer sound continue through a louder one that masks it. It allows us to resist the masking effects of loud sounds upon weaker ones in mixtures (for example, in musical performances). Later we shall see how some rules of counterpoint in polyphonic music harness this heuristic for an artistic use.

The memory that the old-plus-new heuristic can use is fairly complex and is not just a record of a steady-state spectrum. For example, it can be used to extract a gliding pure tone that was heard earlier from a gliding complex spectrum that is currently being received. Its manner of activity is not simply that of a filter, because it makes some of its decisions after the complex mixture has gone by. For example, the continuity illusion is enhanced if the softer sound appears again at the termination of the louder one.

The relative intensity of various partials in a sound also affects the partitioning. A partial that is much louder than its nearby frequency

neighbors is more likely to be heard as a separate sound. The auditory system also appears to treat as more "normal" a spectrum in which the higher partials are less intense. When this pattern is violated, the higher harmonics are easier to extract. Both these phenomena may be interpretable via long-known facts about the spectral spread of masking.

We have also seen that the harmonic relations in a spectrum are important. Partials that are in a common harmonic series are more likely to be treated as the spectrum of a single sound. This also holds when more than one harmonic series is present in the same spectrum; the spectrum is partitioned into more than one subset by means of the harmonicity principle. The separate perception of the spectral components of each subset is suppressed in favor of global qualities of the subset. The harmonicity principle in auditory scene analysis seems to be related to the pitch-extraction mechanism, but not to be equivalent to it.

Heuristics based on the Gestalt principle of "common fate" seem to be used as well. When different partials in the spectrum undergo the same change at the same time, they are bound together into a common perceptual unit and segregated from partials whose time-varying behavior is different. This principle applies both to changes in intensity and changes in frequency.

Spatial factors are also used. It appears that the auditory system is capable of estimating a separate spatial origin for different frequency bands in the spectrum and then grouping those bands into sets on the basis of a common spatial origin. Spatial origin is a strong cue for partitioning the spectrum but not an overwhelming one. It can be overcome by other factors. Indeed, even the spatial estimates themselves can be altered when in conflict with other cues; for example, other cues may tell the auditory system that two parts of the spectrum should be treated as parts of the same sound even though they yield different spatial estimates.

The cues for spectral grouping compete and cooperate until the scene-analysis system converges on the best estimate of how many sounds there are, where they are, and what their global properties are.

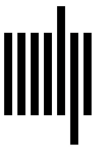
Auditory Scene Analysis

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