#3234

2

Auditory scene analysis: hearing in complex environments

Albert S. Bregman

2.0 INTRODUCTION

I would like to introduce an approach to auditory perception that is concerned with quite different questions from those asked by traditional psychophysics. The latter is occupied with such questions as these: What is the minimum amount of energy that can be sensed by the auditory system? How far apart do the frequencies of two pure tones have to be in order to be distinguished when they are played either sequentially or at the same time? How does the experienced loudness of a tone grow as its physical intensity is increased? How do differences between the acoustic patterns registered at each ear tell us where the source of sound is?

The development of artificial intelligence in recent years has made us aware that even if the above questions were answered with certainty, we would still understand very little about how the auditory system worked. A useful exercise that can help us to see the problems is first to imagine the auditory system as being possessed by a robot, and then to ask what its hearing capacities (as presently conceived) would do for it.

Let us imagine that we equipped our robot with a pair of sensing devices as ears and embodied, in these devices, all the properties that had been found to be true of human audition. The robot would still have great difficulty using the information that it received. Its most difficult problem would be in dealing with mixtures of sounds. Its record of any incoming signal would represent the sum of all sound-producing sources that had been simultaneously active at the time of the recording. Suppose that our robot had a definition in its memory of the sound of a voice saying a particular word. It still might not be able to recognize the word because the presence of other sounds had created a situation in which no segment of the recording matched the definition closely enough. Even worse, it might mistakenly hear some accidental product of the mixture of two voices as a word.

Thinking in sound: the cognitive psychology of human audition, ed. S. McAdams and E. Bigand. Oxford University Press, 1993, pp. 10–36.

Here is another example of the difficulty that it might have. Psychophysicists tell us that the intensities of the sounds at the two ears can be used as a clue for the spatial position of a sound source. But how does the robot know, when it is comparing the intensities at its two 'ears', that it is comparing the energy derived from only one sound source? If there are two, each in a different place, the simple strategy of comparing intensities at the two ears will no longer work. It must do a separate comparison of the intensities derived from each source. How is it to know how much energy came from each source to each ear?

To recognize the component sounds that have been added together to form the mixture that reaches our ears, the auditory system must somehow create individual descriptions that are based on only those components of the sound that have arisen from the same environmental event. The process by which it does this has been called 'auditory scene analysis' (Bregman 1990).

The term 'scene analysis' was first used by researchers in computer vision to refer to how a computer might solve the following problem. In a photograph of a scene of normal complexity, the visible parts of a single object are often discontinuous because the camera's view of the object has been interrupted by the presence of another object that lies between it and the camera (e.g. Guzman 1969). Scene analysis is the name given to the strategy by which the computer attempts to put together all the visible properties—edges, surface textures, colours, distances, and so on—that belong to the same object. Only then can the correct global shape and properties of that object be determined. By analogy, *auditory* scene analysis is the process whereby all the auditory evidence that comes, over time, from a single environmental source is put together as a perceptual unit. This chapter will describe the methods that the auditory system employs and some of the research that has discovered them.

2.1 SCENE ANALYSIS IN AUDITION

When you describe the problem of mixtures to most people, they are inclined to say that they solve it simply 'by paying attention to one of the sounds at a time'. In saying this, they imply that the parts of the same sound are somehow a coherent bundle that can be selected by the process of attention. However, we must remember that the only thing received by the ear is a pattern formed by pressure changes over time, and if we look at a graph of the waveform of a mixture of sounds, there is nothing obvious in it that labels the sound as a mixture or tells you how to take it apart.

We know that the first stage in the analysis of sounds by the human auditory system takes place in the cochlea. There the sound is decomposed

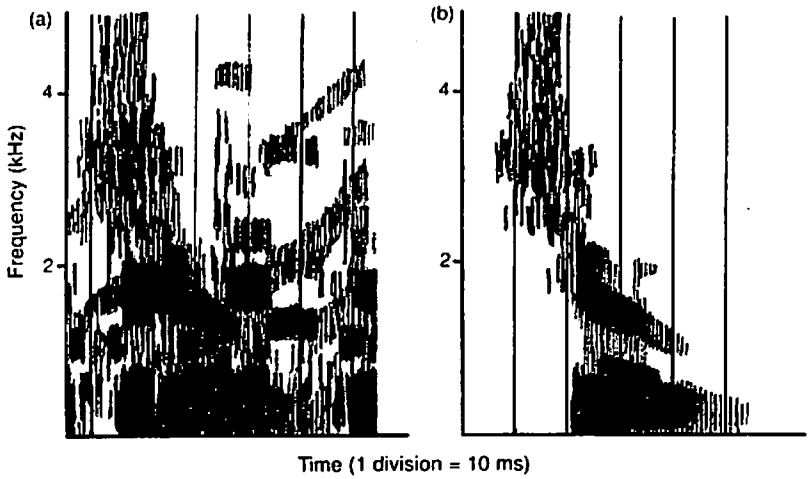

Fig. 2.1 Spectrogram of (a) a mixture of sounds and (b) one of the components of the mixture, the word 'shoe'.

into separate neural patterns that approximately represent the different frequencies in the signal (see Moore and Patterson (1986), for discussions on the limits of the ear's ability to separate very closely spaced frequencies). Decomposition into component frequencies is a technique also used by scientists in trying to understand hearing. The results are displayed in a spectrogram, a picture that shows time on the x axis and frequency on the y axis. The darkness at any point on the picture shows the intensity of the sound at a particular time and frequency. We can appreciate the limitations in the usefulness of the ear's frequency-based decomposition of the signal by considering the information shown in a spectrogram.

The one shown in Fig. 2.1(a) shows a mixture of sounds. The reader might think that the problem of finding the individual sources from mixtures might be solved immediately by using such a picture. This could indeed be done if the sources were steady, pure tones that were well separated in frequency. Then each horizontal streak on the picture would represent a separate environmental sound, persisting for some period of time. Figure 2.1(a), however, represents a more natural case in which each environmental sound has many frequency components and these are not constant over time. (It represents a mixture of a man saying 'shoe', another singing to himself, and an unrelated piece played by instruments in the background). A listener can easily hear the word 'shoe' in the recording from which this spectrogram was made. Figure 2.1(b) shows the word spoken in isolation. This is what must be extracted from the mixture. We can do it visually to some extent after seeing the isolated pattern, but listeners can do it even without being told which word is

embedded in the mixture. The problem they face can be shown with reference to the spectrogram. We can see that in the mixture, the streaks that represent the components of the word are interlaced with and cross over those representing the components of the other sounds. Even a single streak can represent the sum of two or more components of the same or almost-the-same frequency that have been derived from different sounds. So separating the frequencies in the nervous system and laying them out over time, as illustrated in Fig. 2.1(a) does not, in itself, provide coherent bundles that can be selected by attention. The separate sound sources that listeners describe are not given in any simple way in the spectral decomposition of the signal. The problem of finding them is the job of auditory scene analysis.

It appears to me that there are three processes occurring in the human listener that serve to decompose auditory mixtures. One is the activation of learned schemas in a purely automatic way. It is a common observation that occasionally, people imagine they hear their names spoken in a noisy environment, for example a city street corner. Apparently, a chance co-occurrence of sounds can activate the mental schema that represents the sound of one's name. This hypersensitivity and automatic activation presumably occurs because people so frequently hear their names spoken that its schema is in a highly potentiated state. Whenever the incoming sound matches the schema's acoustic definition in even an approximate way, it becomes active. This method of analysis might penetrate some mixtures as long as the sound pattern recognized by the schema was not totally distorted.

A second process that can decompose mixtures is the use of schemas in a voluntary way. An example occurs when we are intent on trying to hear whether our name is being called out by a person announcing the next appointment in a busy office. The experience of 'trying' is an indication that voluntary attention is involved. Notice, however, that the schema for our name is also involved, though its operation is not automatic in this case. In general, whenever we are listening for a specific sound or class of sounds, some criterion for recognizing the targets must be employed. I call this criterion a schema because it is a mental representation of a particular set of characteristics.

Both of these methods, automatic and voluntary recognition, require that schemas (knowledge of the structure of particular sounds or sound classes that are important to us) have *already been formed* by prior listening. If these schema-based methods were the only ones for decomposing mixtures, it would be hard for us to form schemas for important sounds in the first place unless we frequently encountered these sounds in isolation. It would be valuable, therefore, for us to have general methods for partitioning an incoming mixture of sound into separate acoustic sources that could be used prior to any specific knowledge of the important sounds of our environment.

It appears that we do have such methods. I have referred to them collectively by the term 'primitive auditory scene analysis' (Bregman 1990, p. 38). By calling them primitive, I mean that instead of depending on knowledge of specific types of sound, such as voices, musical instruments, or machines, they depend on general acoustic properties that can be used for decomposing all types of mixture.

2.1.1 Using general acoustic regularities

There are certain relations between the acoustic properties of a sound that can be useful in solving the scene analysis problem even if the listener is not familiar with the signal. These relations are the result of general properties of sound-producing events and are not specific to voices, music, animal sounds, or any other individual class of sounds.

An example of a general property is the fact that many sounds of our environment are harmonic. That is, their frequency components are all multiples (approximately) of a fundamental frequency. This arises from the fact that when force is applied to certain types of bodies, they go into oscillatory motion. Their sub-parts also go into oscillation. Sub-parts forming halves, thirds, quarters (and so on) of the whole body go into oscillations whose frequencies are respectively twice, three times, four times (and so on) the frequency of the oscillations of the whole body. Examples of such bodies are the ones that produce animal vocalizations (including the human vocal cords), the strings in musical instruments, and parts of many machines.

The auditory system can employ a strategy that exploits this regularity. For example, what should it conclude about a signal, when it discovers from an analysis of its frequency components over a short span of time, that all the pure-tone components are multiples of a common fundamental? Are there many sounds present or only one? What are the odds in favour of a hypothesis that there are a number of sounds present, each with a different set of harmonics, all of which happen, by chance, to be multiples of a common fundamental? This would be a highly improbable accident. A better bet is that it is hearing only one acoustic event that is generating harmonics. Here is another example: What should the auditory system bet if it detects two subsets of components, the members of each of them being multiples of a different fundamental? Obviously, the best bet in this case is that the signal is a mixture of two harmonic sounds. The use of harmonicity is a case, then, where a general regularity of the world can be used to decide on the number of sounds being heard.

We find, however, that no single regularity can be trusted all the way. After all, it is not impossible that the harmonics of two sounds might briefly enter into an apparently regular relation with one another. But we would find, in all likelihood, that they were coming from different places

in space or that they started and stopped at different times. Since the frequency components from a single acoustic event tend to come from the same place and to start and stop at roughly the same time, differences in these properties would protect us against accepting accidentally harmonic components as parts of a single sound. We therefore need to exploit many regularities at the same time if we are to come to the right answer.

Shepard (1981) has argued that because animals have evolved in a world that contains regularities, it is likely that their perceptual systems have evolved modes of operation that take advantage of them. That is, as animals evolved, the properties of their perceptual systems became tuned to the regularities of the physical world. The resulting match is referred to as 'psychophysical complementarity.' If Shepard is right, a good strategy for finding the laws of auditory organization would be to try to discover relations among the components of the incoming sound that are frequently present when parts of the sound have been created by different environmental events. Then we can do experiments to find out whether the human auditory system exploits these relations to decompose mixtures.

2.2 EFFECTS OF AUDITORY SCENE ANALYSIS ON PERCEPTION

The organization of sensory evidence which is carried out by auditory scene analysis, affects many aspects of auditory perception that we do not usually believe to be related to perceptual organization. Organization is often viewed as a grouping of raw auditory properties *after* they have been created by perception. However, even such apparently raw properties as loudness can be affected by organization.

Here is an example described by Warren (1982) as 'homophonic continuity' (see also Warren *et al.* 1972; Warren 1984). Suppose we are presented with a steady sound which first holds a fixed intensity for a few seconds, then suddenly becomes more intense, holds this new intensity briefly, and then returns to the original level and continues for some time. If the period of greater intensity is short, we hear it not as a change in the original signal, but as a second sound, identical in properties to the original one, joining it, and then disappearing. The first signal is heard as continuing, unchanged in loudness, behind the added one. The intensity, when it is at the higher level, is interpreted not as coming from a single loud sound, but as a mixture of two softer sounds. In other words, the intensity information is shared out between two perceived events. On the other hand, in experimenting with this stimulus, I have noticed that if the rise and fall in intensity are not sudden, then

the two-sound interpretation does not occur (Bregman 1991). Instead, the original sound is heard as changing in loudness, and the full intensity at the high level is used to derive the perceived loudness of a single sound. To summarize, organizational processes will decide whether we hear one loud sound or two softer ones.

We can create a variation of this example to show how perceptual organization can affect even a sound's perceived spatial position (Bregman 1991). Let us present the signal I have just described, but this time to our right ear only. At the same time, we present to the left ear a sound that is identical to our original signal in all respects, except that it always remains at the lower intensity. During the initial phase, when both the left and right ear are receiving the same lower-intensity signal, we hear a single, steady sound in the middle. Then when the right-ear signal suddenly rises in intensity, stays briefly at this level, and then returns, we perceive the original sound continuing in the middle, accompanied briefly by a sound at our extreme right. With this procedure two sounds are heard, each with its own position.

Suppose, instead, we raise and lower the intensity at the right ear less suddenly. In this case, we hear the sound move over to the right and then back to the centre, our spatial perception following the changing balance of intensities at the two ears. We perceive only a single sound with a single location.

Regardless of whether the change is sudden or gradual, it leads to the same intensity balance between the two ears during the middle phase (when the sound is more intense at one ear). If perceived location were unaffected by the history of the event, we would hear the same thing in the two cases. But we do not. This tells us that when we perceive a location, it is not the location of 'sound' in the abstract, but of a particular sound. Depending on how many sounds are created by perceptual organization, the set of perceived locations will be different. The simple rules for spatial perception that classical psychophysics has discovered by testing listeners in simple, quiet environments cannot be applied without modification in acoustically complex ones.

Auditory scene analysis can also affect whether we hear one tone with a rich timbre or a number of tones with purer timbres (Bregman and Pinker 1978). It can affect the perceived identity of a speech sound by segregating acoustic components that might otherwise be part of it, and allocating them to another perceived event (Darwin 1984; Ciocca and Bregman 1989). It can determine whether two melodies, presented as a mixture, are heard individually or whether a new emergent melody, formed of all the notes, is heard instead (Dowling 1973). It can determine whether the instruments of an ensemble are heard with their individual properties or blend to form a global timbre (Bregman 1990, Ch. 5). In short, it affects every auditory experience in natural environments.

Although I have surveyed much of the research on auditory scene analysis at length in a separate volume (Bregman 1990), I will try here to convey a basic understanding of the approach. The presentation will be organized around the different environmental regularities exploited by the auditory system.

> ***Regularity 1.*** *Unrelated sounds seldom start or stop at exactly the same time.*

Acoustic components derived from independent environmental events tend not to start and stop at the same time. Instead, one is likely to be active already at the moment that another begins. The auditory system exploits this general truth when it uses what I have called the 'old-plus-new' strategy: when a spectrum suddenly becomes more complex but, as far as the nervous system can tell, it still contains the same frequency components as before, it is interpreted as a continuation of an old signal with a new one joining it. The old sound continues to be heard for a short time along with the new one. If the spectrum becomes simple again, so that it contains only the components of the old sound, this strengthens the perception of the old one as having been present all the time. The perceived qualities of the added sound are derived from the components of the complex spectrum that are left over after the components of the earlier simpler spectrum are subtracted out.

The various phenomena of perceived continuity (see Warren 1984) are derived from this strategy. In the best known example, a long pure tone is alternated with a short, louder noise burst that stimulates neural channels that include the one stimulated by the tone. A listener will hear the tone as continuing through the noise—it seems to be present all the time. Rather than hearing the tone turn suddenly into a noise burst, the auditory system determines that the neural activity occurring during the noise burst is consistent with the continuation of the tone. This allows it to generate a percept of the tone continuing behind the noise. Probably the rejection of the alternative interpretation, a tone turning into a noise, is due to the auditory system's exploitation of another environmental regularity that we will discuss later: when properties are derived from a single continuing sound, they tend not to change suddenly.

In the preceding example of continuity it is hard to determine whether the part of the neural activity that was interpreted as the continuing tone was allowed to contribute to the perception of the noise. A broad band of noise sounds very much the same whether or not the narrow band of frequencies allotted to the tone has been subtracted from it. A better stimulus for seeing how the total information from the noise is shared out is a signal in which a band of noise, say 0 to 1 kHz ('kHz' means

'thousands of cycles per second'), alternates with a wider band (say 0 to 2 kHz), with the narrower band of noise longer in duration than the wider one. The stimulus is generated so that the components shared by the two sounds (i.e. those in the 0–1 kHz range) are equal in intensity in the two sounds. In this demonstration, a listener will hear the narrow-band noise as continuing through the wide-band noise (Warren 1982, Ch. 2). In addition, it is the listener's impression that the short intermittent sound lacks the components that have been used by the auditory system to create the percept of the continuing narrow-band noise. The contribution of the 0–1 kHz components has been removed and the 'added' noise is experienced as having the same quality as a 1–2 kHz noise. The spectrum of the wider-band 0–2 kHz noise has been divided up, providing separate components for the continuing and the added sounds.

An example of the old-plus-new strategy in the field of speech perception is found in an experiment by Darwin (1984). Each vowel in a language can be distinguished from all the others by the positions of a number of formants (intensity peaks) in its spectrum. Each formant consists of an augmented intensity of a number of harmonics. Darwin lengthened one of the harmonics in the lowest formant of a vowel so that it started before the vowel and continued throughout its duration. In this stimulus pattern, the tone is the 'old' sound and the vowel is the 'new' one. The effect was to cause the vowel to sound more like a different one. This occurred because when the harmonic was heard alone, it was assigned an identity as a separate sound. Then when it was joined by the rest of the vowel, the total neural stimulation activated by the vowel was partitioned. A part was interpreted as the continuing (old) tone. Only the remainder, with this part removed, was interpreted as the vowel. The remaining sensory evidence pointed to a peak in the spectrum at a different place from that shown by the whole body of evidence, and, in doing so, changed the identity of the perceived vowel.

> ***Regularity 2.*** *Gradualness of change.*
>
> *(a) A single sound tends to change its properties smoothly and slowly.*
>
> *(b) A sequence of sounds from the same source tends to change its properties slowly.*

The environmental regularity concerning the sequential resemblance of sounds from the same source tends to take two forms. One of them relates to a *single sound-producing event* that extends over time. Examples (at different scales of time) are a single violin note, a single syllable being spoken, the roar of a lion, or the continuous noise of a motor. The sound

coming from such an event tends to change continuously rather than abruptly in its properties as the event unfolds.

A second form of this sequential regularity is concerned with a *succession of sounds* from the same source. Many examples can be cited: a succession of footsteps, a series of chirps in a bird's or a frog's call, a series of pecks of a woodpecker, and so on. Sounds derived in succession from the same acoustic source tend to resemble one another, with only gradual changes in properties between members of the series.

Sometimes it is not the event that is occurring repeatedly, but our auditory access to it. In a mixture of sounds, each of which is waxing and waning in intensity, the auditory system may obtain a succession of exposures to the properties of one of the sounds relatively unmixed with properties of the others. The spectral samples caught in these successive 'glimpses' of the sound are likely to resemble one another.

An important fact about sequential resemblance is that sounds do, in fact, change. However, since the changes tend to be gradual, samples from the same acoustic event that are taken closer together in time tend to resemble each other more closely.

These facts about sequential change in the sounds of the world are encapsulated in two rules that seem to be followed by the auditory system. The first rule that is based on these regularities is the 'sudden-change rule'. The auditory system will treat a sudden change of properties as the onset of a new event. Examples of this were given earlier in this chapter. We had the example of homophonic continuity as well as the example in which one ear received an abrupt rise in intensity while the other ear's sound remained steady. In both cases, the suddenness of the change triggered the interpretation of an added sound. This set the stage for the old-plus-new strategy to decompose the spectrum into the components belonging to the old sound and those of the added one.

The use of this strategy requires a definition of suddenness. However, there is no clear boundary between gradual and sudden changes. This is illustrated in a recent unpublished study by Jean Kim and myself. We played a rapid sequence of four one-second tones to subjects and asked them to judge their order of onset. They were highly overlapped in time—for example each successive tone could begin as little as 0.15 s after the previous one. The suddenness of the rise in intensity of each tone, as it was turned on, helped the listener to decompose it from the mixture. When the onsets took 0.04 s, the tones sounded distinct and the listener could judge their order fairly well, but when it took 0.64 s, the tones were all blended together in an impenetrable mush of sound and it was almost impossible to judge their order. However, the change from sudden to gradual was not 'all or nothing' in nature. An intermediate onset time, 0.16 s, gave an intermediate result. The suddenness of spectral change

works like every other clue for scene analysis. The stronger it is, the more it affects the grouping.

The range of durations (0.04 to 0.64 s) that I have described as embodying a wide range of suddenness may be a valid range only for changes in intensity. The values that count as sudden and gradual for changes in other features of sound, such as spatial location, timbre, or pitch, have not yet been pinned down quantitatively.

Sometimes the auditory system will not be able to find a close enough match in the spectra before and after the sudden change to support the old-plus-new interpretation. Therefore, instead of interpreting the change as a second sound being added to the mixture, or as a change in the qualities of a single continuing sound, it may conclude that a second sound has replaced the first. An experiment by Darwin and Bethell-Fox (1977) found an example of this in speech perception. First it is important to understand that the fundamental frequency and the frequencies of the formants can be varied independently of one another. The fundamental is the frequency of which all the components of the spectrum are multiples (harmonics). The formants are those parts of the spectrum in which the harmonics have been strengthened in intensity. The researchers synthesized a sample of continuous speech consisting of smooth formant transitions (gradual changes in the positions of spectral peaks) back and forth between two vowels. The vowels were represented by frequencies of the formants that were held steady for 0.06 s. In the middle of each transition, they suddenly changed the speaker's fundamental frequency, so that one vowel was always on a lower frequency, 101 Hz, and the other on a higher one, 178 Hz ('Hz' means 'cycles per second'). This caused a segregation of two apparently different talkers, one speaking in a lower pitch and the other in a higher. Furthermore, even though there was a continuous change in the formant frequencies, the parts of the transition before and after the abrupt change in pitch tended to be perceptually isolated from one another. Each voice seemed to be silent during the period in which the other was talking. Furthermore, the syllables that were constructed in perception tended *not* to be based on formant patterns that bridged the change in the fundamental. It was as if the auditory system did not want to make a mistake by creating syllables out of relations between the sounds from two different talkers.

A second rule that exploits the probable sequential resemblance of sounds from the same source is the 'grouping by similarity' rule. We do not yet know all the ways in which sounds can be similar. However, similarities in the frequencies of pure tones, in spatial location, and in the spectral content and fundamental frequency of complex tones has been shown to affect grouping (Bregman 1990, Ch. 2). We are also unsure about the measurement of similarity. For the frequencies of pure tones,

the difference that affects grouping appears to be the difference between the logarithms of their frequencies, but for other dimensions, no one has yet tried to discover appropriate quantitative measures of the degree of similarity. It has been directed, instead, towards demonstrating that such similarities actually have an effect on grouping, using common sense assessments of similarity.

In some sense, we have already discussed similarity. We might argue that discontinuity, mentioned earlier, is merely the absence of sequential similarity. However, in mentioning the similarity rule separately, I am thinking of cases in which the sounds occur in discrete samples or episodes, such as a succession of footsteps or a succession of glimpses of one sound in a mixture whenever the others drop in intensity. These successive samples of sound are modelled in the laboratory by successions of tones, bursts of noise, tonal glides, and so on.

The effect of the rule is to take sounds that have similar properties, to link them together perceptually into groups, and to segregate these groups from one another. Each linked group is considered by the auditory system to have come from a distinct environmental source. As a result, pattern-recognition processes tend to treat the grouped sounds as the 'package' of evidence within which to look for familiar patterns.

The phenomenon of auditory streaming exemplifies the working of this rule (Bregman and Campbell 1971; van Noorden 1975, 1982). Suppose we start with two sets of tones, one of high tones and the other of low ones, in which there is a considerable frequency separation between the high and low sets, but where the frequency differences between the tones within each set are small. Suppose we then interleave the high (H) and low (L) tones together (e.g. HLHLHL . . .) to make a long sequence, and play it to listeners. Recall that I said that in the world at large, the closer in time two samples from the same acoustic source are, the more similar their properties are likely to be. The 'grouping by proximity' strategy takes this regularity into account. As long as the succession of tones is slow (i.e. the samples are far apart in time), the auditory system is willing to accept the entire set of tones as coming from a single source. However, when the sequence is speeded up, two perceptual 'packages', or auditory streams, are formed—the listener hears the sequence as if two distinct sources of sound were active at roughly the same time. Patterns formed by the high and low tones taken in combination tend to disappear in favour of patterns formed by the high tones alone and the low ones alone. For example, two melodies will be heard, one involving the high notes and the other the low ones. Also if the original mixed sequence is irregular, different temporal patterns will be heard in the high and low ranges. For example, the following sequence of high and low tones

...H H L H L L H L H H L L H L...

will break into the streams

...H H - H - - H - H H - - H -... and
...- - L - L L - L - - L L - L...

In these letter-diagrams, the dashes represent silences that appear in each stream as a result of the absence of the tones that have been placed into the other stream by the perceptual grouping. So, temporal patterns, including rhythms (in repetitive sequences), can be created or altered by the segregation into auditory streams.

The grouping is sensitive to the degree of difference between the frequencies of the high and low tones, the segregation becoming stronger as the separation in frequency becomes greater. This frequency difference can trade off against the speed of the sequence. A sequence in which the frequency difference is smaller must be sped up more before it will be segregated (van Noorden 1975).

The sensitivity of segregation to the speed of the sequence can be viewed as a sensitivity to the rate of change between high and low tones. When changes occur more rapidly, the sequence is less likely to be treated as issuing from the same environmental source. This property of the perceptual process is justified by the environmental regularity that, in a series of sounds from the *same* acoustic source, the properties tend to change only slowly.

An important fact about the segregation of the tones of two classes (at least when the classes are defined by frequency range) is that it is cumulative. Even at a high speed, the first few high and low tones are usually integrated into a single stream. As the auditory system continues to hear two populations of tones, segregation occurs. It appears that the segregative tendency builds up for at least four seconds. Just as the tendency to segregate takes some time to build up, it also takes at least four seconds to dissipate. Even if the sequence stops for a couple of seconds, some segregative tendency will remain, and if the sequence starts again, it will be segregated more quickly the second time (Bregman 1978). This auditory principle seems to correspond to the environmental fact that the best predictor of a sound occurring is the fact of having just heard it a moment earlier. Apparently the auditory system keeps an 'open slot' available for the return of the sound.

It appears that even simple perceptual judgments are affected by the segregation of components in different frequency ranges. The exact temporal separation between two tones, one higher in frequency than the other, becomes more difficult to judge as the frequency separation is increased (see Bregman 1990, pp. 159–63 for an account of this research).

2.3 MULTIPLE BASES FOR SEGREGATION

The frequencies of the tones are not the only differences that affect their sequential organization. Tones can also be grouped according to similarities in their spatial positions (described by Bregman 1990, pp. 73–83). In addition, if the sounds are complex, other factors can play a role. In tones formed of many harmonics, as voices and musical tones are, the grouping can depend on similarities in their fundamental frequencies and also in their timbres, if we define timbre by the relative strengths of harmonics in their spectra (Singh 1987; Bregman *et al.* 1990*a*). Tones (sounds whose waveforms repeat cyclically) will often segregate from noises (Dannenbring and Bregman 1976). Tones that glide from one frequency to another will connect together sequentially better when they have the same slopes and are in the same frequency range (Steiger and Bregman 1981).

These various types of difference compete and cooperate with one another in determining grouping. If different factors promote contradictory groupings of the sounds, the winner will be the grouping with the most factors favouring it or the grouping that is favoured by the factors that the auditory system prefers to use (Bregman 1990, pp. 165–71, 218, 335).

This brings up the point that not all acoustic differences are equally important in determining grouping. Hartmann and Johnson (1991) asked listeners to recognize pairs of melodies whose notes had been interleaved to form a new sequence of tones. In each sequence, the notes of the two melodies were made to differ from one another on a single acoustic characteristic. The most effective ways to help people to segregate the two melodies were to send them to different ears, to shift the notes of one melody up by an octave, or to introduce a timbre difference between them (pure sine-wave tones versus rich complex tones). Virtually no improvement of segregation was obtained when the difference involved the attack and decay times of the notes of the two sets or when the rhythm of the overall sequence was altered by shifting one set of notes so that they did not fall exactly half way between the notes of the other, or when noise was added to the notes of one melody.

When we observe, in an experiment, that segregation is more strongly affected by some acoustic factors than others, we still do not know whether the observed differences arose because of their effects on the primitive scene-analysing mechanism or because of the activity of more sophisticated mechanisms that use learning or attention. Evidence that there is more than one mechanism comes from the different effects that speed has, depending on the intentions of the subject. Van Noorden (1975) played his listeners a sequence in which tones in two different frequency regions alternated, and asked them about the perceptual segregation of the tones into high and low streams. He varied the frequency separation

between the high and low tones and the speed of the sequence. The results he obtained depended on what he asked the listeners to do. When they were asked to hold the high and low tones *together* in a single stream as much as possible, he found a trade-off between frequency separation and speed. As the speed became higher, the tones had to be closer together in frequency before they could be integrated. However, there was no such trade-off when they were asked to try to *segregate* the high tones from the low ones as much as possible. As long as there was a minimum frequency separation of a few semitones, the sequence could be segregated equally well at any speed. To me this difference indicates that there is more than one mechanism of segregation. When listeners are trying to *integrate* the sequence, the segregation is involuntary, acting in *opposition* to their intentions. This sort of segregation is probably due to primitive scene analysis mechanisms. On the other hand, the segregation that occurs when listeners are trying to segregate the sounds, i.e. when the segregation is *consistent* with their intentions, is the product of a selection process carried out by attention (Bregman 1990, Ch. 4). The existence of two mechanisms explains another related finding. When we vary an acoustic difference among some sounds in a sequence, two aspects of the results are not always consistent with one another. For example, differences between notes on timbre may *assist* the listeners strongly when they are trying to segregate the sequence on that basis. However, those same timbre differences may *not* strongly oppose the grouping when listeners are trying to segregate the sequence on the basis of another factor, pitch differences (Bregman *et al.* 1990*a*). If only one process of segregation existed, you would expect that when it supported your intentions, it should give symmetrically opposite effects to those obtained when it opposed your intentions.

If more than one mechanism can be involved, experiments that study the role of acoustic differences in segregation must attempt to ensure that the variation among the results of different experimental conditions is due to only one of these mechanisms. We have tried in our own research at McGill University to engage the primitive mechanism by always giving subjects tasks that require them to hold the sequence together. When they cannot do so, the negative influence is attributed to a primitive mechanism.

Consider the Hartmann and Johnson experiment mentioned earlier (p. 23). The listeners were required to pull out a sub-sequence of sounds (a melody) whose tones were marked by a common property. Therefore, they could have set their attentional mechanisms to select tones with that property. The difficulty they had in using some of these properties could have come because either the primitive mechanism or the attentional mechanism made little use of them. It is likely, since all differences were discriminable, that the results were influenced primarily by primitive

grouping. However, a purer test of primitive grouping would be to control the listeners' attentional strategies by asking them to try to use the same acoustic difference (e.g. attack time) on *every* test, but to oppose their intentions by introducing other differences that could be employed by primitive scene analysis to form groupings that opposed the intended ones. The effectiveness of this opposition would reveal the effectiveness with which primitive grouping used that property.

The readiness of the auditory system to group similar sounds when they occur in a sequence is the basis for a slightly different version of the 'old-plus-new' strategy from the one I described earlier. In the examples I gave before, one of the components of a mixture started ahead of the mixture as a whole and continued, without a break, into it. The clear glimpse that the auditory system obtained of the earlier-starting sound allowed it to be factored out of the mixture. A similar, but weaker, effect occurs when two sounds are used—a simpler one, and a more complex one that contains the simpler one as part of it, but there is a break between them. They are rapidly alternated, with brief silences between them, in a repeating cycle. Under favourable circumstances, the listener will experience the pure tone twice on each cycle, once when it is presented in isolation and once when the complex tone occurs. This means that the isolated pure tone has captured the corresponding frequency out of the complex tone into a pure-tone stream. As a result, the listener will hear the pure tone twice per cycle (the second occurrence is the component captured out of the complex tone). The remainder of the complex tone will form a second sound that seems to occur only once per cycle.

This extraction of an earlier-heard sound out of a complex one was described earlier as the old-plus-new strategy. The only difference in the present case is that there is a silence separating the isolated tone from its counterpart in the later mixture of components that comprises the complex tone. However, because the isolated tone (the captor) and its 'target' inside the complex tone are not continuous, this stimulus pattern allows us to manipulate the frequency proximity between them. If the captor and target are pure tones, then increasing the frequency difference between them reduces the capturing. With sufficient separation, the second occurrence of a pure tone on each cycle can no longer be heard (Bregman and Pinker 1978).

This influence of frequency proximity in the capturing of a tone from a complex spectrum is a direct counterpart to its influence in the sequential streaming of tones that occur in different frequency ranges. The two cases can be described in the same terms: an earlier tone (or stream of similar tones) tries to link itself to a newly arriving tone (*or a part of it*) in proportion to the proximity of the new component to those already in the stream. In both cases, there is a sequential grouping by proximity.

2.4 DIFFERENCES IN SPATIAL LOCATION

Regularity 2 (gradualness of change) applies to spatial location as well as to the other factors already mentioned. Sounds that are created by the same event typically come from the same position in space, or from a location that is changing slowly. Correspondingly, there is a scene-analysis principle that groups sounds that come from the same spatial location. Bregman (1990) describes cases in which this grouping occurs when the sounds appear sequentially (pp. 75–83), or when they are presented at the same time (pp. 293–312). An example of the latter can be created by sending the upper- and lower-frequency components of a speech sound to separate ears. In certain conditions, a separate sound will be heard on each side of the head (e.g. Cutting 1976). This segregation of speech components, while audible, often does not prevent their use together to form a speech sound. This freedom of speech perception from compulsion by primitive scene analysis resembles what happens in the case of segregation by harmonic relations. I discuss the latter further on, together with the reasons why it may occur (see also Bregman 1990, Ch. 7).

While the use of spatial location is central to many attempts by engineers to programme computers to segregate the sound of a person speaking from other co-occurring sounds, humans do not seem to depend so heavily on this cue. They do use spatial location, but when it competes against a sequential grouping based on frequency differences, it typically loses (e.g. Smith *et al.* 1982).

When two steady sounds are played at the same time in different spatial locations, the listener's ability to derive a separate estimate for the location is not very precise (Divenyi and Oliver 1989). Yet I believe that there *is* a role for spatial differences in auditory scene analysis. I would guess that, in general, they play a facilitating role, strongly enhancing segregation based on other factors such as asynchrony, or differences in frequency or timbre. They would probably greatly facilitate the following of sounds into mixtures when the earlier sound and the added ones were at different locations.

Why should the human not give absolute priority to the spatial cue? I can imagine both a reason based on the environmental information available to scene analysis and one based on the physiology of audition. First, consider the environmental argument. The gradualness of change of the location of an event in our environment should be just as strong a regularity as gradualness of change of frequency. While this may well be true, the physics of sound makes the evidence about location less reliable. We use differences in the sound received at our two ears or the way it is filtered by our outer ears to derive the direction from which it was coming. However, sounds can bounce around corners in the environment, or reflect off walls near one of our ears, or an obstructing object

can move close to one of our ears, attenuating the sound. These phenomena cause the information about location to become incorrect. Such events cannot, however, change the fundamental frequency of a sound or its internal harmonic relations or add a frequency that was not there before; therefore we should not be surprised that fundamental frequency, harmonic relations, and frequency composition are used more strongly than spatial location. Another possible reason for the conservative use of the location cue may be found in the physiology of hearing. To segregate sounds based on their locations, the auditory system must calculate which frequency components have come from the same spatial location; that is, it must assign a separate location estimate to each one. This may be difficult to do when the components are densely packed in the spectrum.

> ***Regularity 3.*** *When a body vibrates with a repetitive period, its vibrations give rise to an acoustic pattern in which the frequency components are multiples of a common fundamental.*

I described this regularity earlier and outlined the corresponding strategy that could be employed to group the partials (frequency components) in a spectrum. Let me now describe a few observations and experiments which show that people do indeed use such a strategy. The most obvious observation is that people can hear the individual pitches of two complex tones played at the same time, even if they start and stop together. We have reason to believe that in a complex tone, many of its partials contribute to its pitch (see Moore 1989, Ch. 5). But, to derive each pitch correctly, the auditory system must be including only the partials of one of the tones and excluding the partials of the other one. In the present example, this could only be done by using the harmonic relations between the partials of the same tone. This example shows only that the auditory system can use harmonic relations for deriving two pitches at the same time. It does not prove that harmonic relations can be used to segregate groups of partials for any other purpose. This proof, however, has come from laboratory studies.

Darwin and Gardner (1986) showed that a segregation based on harmonic relations could effect the perception of a vowel. They based their procedure on a finding by Moore *et al.* (1986); so let me diverge a little to describe this finding. These researchers started with a complex tone in which all partials were harmonically related to the same fundamental. Such a tone is perceived as a unified whole, with only a single pitch. They found that when a low harmonic was mistuned from its harmonic value by a sufficient amount, it was heard as a separate tone with a pitch different from that of the complex tone. Accordingly, Darwin and Gardner

synthesized a vowel with a harmonic spectrum and then mistuned one of its harmonics. The perceived identity of the vowel was changed when the harmonic was mistuned by eight per cent. Apparently this occurred because the mistuned partial was removed from the package of evidence defining the vowel, and the remaining evidence pointed to a different vowel. So the segregation based on harmonicity can affect not just the pitch but other qualities of the sound.

None the less, there seems to be a discrepancy between two effects of harmonic relations—effects on the perception of how many sounds are present, and effects on the identification of speech. The discrepancy has been noticed in cases in which there is only a single speech sound in the spectrum. The spectrum is divided into two non-overlapping regions and a different fundamental frequency is used to derive the harmonics of each region. While listeners will segregate the two regions as far as the perception of pitch is concerned (they hear two), they will still integrate the regions to derive the identity of the speech sound (Cutting 1976; Darwin 1981). We encountered this discrepancy earlier in our discussion of the segregation of parts of a speech sound by their spatial locations.

The finding about fundamental frequencies seems to imply that speech perception does not make any use of differences on this factor to segregate mixtures of sounds. Yet this is not true. When *two* complete speech sounds are mixed, with their spectra *overlapping*, their identities are perceived much more clearly when the harmonics of the two are related to two different fundamental frequencies (Brokx and Noteboom 1982; Scheffers 1983). Spatial differences also make it easier to segregate two concurrent voices (Schubert and Schultz 1962). I believe that it is possible to resolve these contradictions by arguing that the grouping created by primitive scene analysis does not lead to separate percepts in a single step (Bregman 1990, Ch. 7). It merely lays constraints of a non-binding nature on a subsequent description-building process. The latter process is also governed by our schemas of speech sounds and other familiar sounds. The interaction of the two sorts of constraint leads to the final percept.

> ***Regularity 4.*** *Many changes that take place in an acoustic event will affect all the components of the resulting sound in the same way and at the same time.*

An example of this regularity can be found in the sound of a man dragging a load by fits and starts along an irregular gravel road. The sound made by the dragging has many frequency components in it, yet the intensities of all these components rise and fall together in synchrony as the load moves irregularly along. A correlated pattern of intensity change occurs across the spectrum. However, the frequency components

arriving from other co-occurring events, such as a car passing or a person talking at the same time, have their own independent patterns of change.

The perceptual rule based on this regularity favours the grouping of components of the spectrum that have the same pattern of intensity variation. Its effects can be observed in the laboratory in a phenomenon called 'comodulation masking release' or CMR (Hall *et al.* 1984; see also review by Moore 1990). In one version of it, a target sound, such as a pure tone, is played together with a noise whose component frequencies fall in a narrow range (a 'narrow band') and are close to those of the masker. If loud enough, this noise (called the 'on-target band') tends to mask or drown out the target noise because noises that have frequencies close to those in a target are effective in masking it. The amplitude of the masker, but not of the target, fluctuates over time. The experiment proceeds as follows. First the on-target masker is made just intense enough that the target cannot be heard. Next, a second narrow band of noise is added to the signal. Its frequencies are too far from that of the target to mask it. This second band is called the 'flanking band'. Surprisingly, if the amplitude fluctuations of the flanking band are synchronized with those of the on-target band, the perception of the previously masked target signal is restored. Apparently the scene-analysis system can group neural activity across frequency ranges when it shows the same pattern of change. The information summed across the ranges concerning the fluctuations allows the two bands to be treated as a single source that is separate from the target tone, which does not have the same fluctuations. This example shows that not only does the auditory system have the power to group components that share the same pattern of fluctuation but that it can segregate them from others that do not share it.

It is not yet clear as to what exact process permits the segregation to take place in the CMR phenomenon. Clearly, in a natural environment, the combined information from two frequency bands would specify the fluctuations in a source event more exactly than either could do alone, since it is likely that the individual bands would contain a contribution from other events that had different fluctuation patterns. In the CMR case, the fluctuation in the on-target band is being altered through its mixture with the target tone, which is not fluctuating.

But how is the improved information about the fluctuation pattern employed to better detect the target? Is the fluctuation pattern subtracted from all spectral bands that show any sign of it? How would the system know how much to subtract? After all, the fluctuating event in the environment would probably not generate equally strong acoustic energy in all frequency bands. It is not clear how the auditory system solves these problems.

Another example of the grouping of components by their fluctuation patterns is found in listening to the human voice. The pulses of air through

the vocal folds of the speaker give rise to a pattern of harmonics. These are spaced apart in frequency, as harmonics always are, by an amount equal to the fundamental frequency. (For example, the harmonics of a 100 Hz fundamental are 100, 200, 300, and so on, all separated by 100 Hz.) When they reach a listener's inner ear, a number of consecutive harmonics can affect the same region on the basilar membrane (see Glossary for explanation). Due to the geometry of the inner ear, this happens increasingly as the frequencies of the harmonics become higher. At higher frequencies, components having a given frequency difference affect places that are closer together on the basilar membrane. When this happens, that region tends to register beats at a rate that is equal to the frequency separation of the harmonics. Because the harmonics are all equally spaced in frequency, this beat rate is the same at different places on the basilar membrane, corresponding to different frequency bands in the incoming signal. This correlated beating in different frequency regions of the auditory system tells it that these regions have been stimulated by the same voice. This can be used as a reason for grouping the evidence from these neural channels. The amplitude fluctuations that are being discussed here are much faster in the case of the voice, even the male voice (mostly between 80 and 200 fluctuations per second), than in the CMR phenomenon (commonly around 10–25 Hz). However, laboratory studies have shown that amplitude modulation around 100 Hz can be used as a basis for fusing the perception of two spectral regions (Bregman *et al.* 1990*b*) and differences of as little as 5 Hz can measurably reduce the fusion (Bregman *et al.* 1985).

2.5 SYNCHRONIZED FREQUENCY CHANGES

When a sound-generating object changes its properties so that its fundamental frequency gets higher or lower, all the partials (pure-tone components) of the sound also change synchronously and in parallel (on logarithmic frequency coordinates). An example of this regularity can be found in the human voice. A voiced sound—a vowel, for example—contains many harmonics. When we raise the pitch of our voice by tightening our vocal folds, the frequencies of all the harmonics rise by a proportionate amount at the same time. The same thing holds true when we move the slide of a trombone.

Suppose that, in a particular listening situation, the auditory system encountered a spectrum in which two subsets of partials had the following properties: (i) the partials within each subset changed their frequencies in parallel, and (ii) the two subsets had independent patterns of frequency change. How should this evidence be used? Obviously it indicates the presence of two independent sounds, each formed of a set of partials.

The auditory system should fuse each subset, i.e. it should use only the partials within a subset to derive a description of a sound. The net effect would be the segregation of two sounds in a mixture.

It has been proposed that the auditory system does exploit this form of regularity. McAdams (1984) found that when a number of pure-tone components, played at the same time, were caused to move through small changes in frequency, the listeners heard fewer sources of sound when their frequencies changed in parallel (on a logarithmic frequency scale) than when they changed in a non-parallel way. This argues that the parallel components were fused. However, in another experiment by McAdams (1989), the auditory system did not seem to be able to use the independent movement of three subsets of components in order to segregate them. The author synthesized three vowels on different fundamental frequencies and mixed them. In all cases he jittered the fundamental frequency of the vowels over time, causing all the harmonics of any individual vowel to change in parallel. In some cases, the pattern of change was the same for all vowels, and in others, each vowel had a different pattern. One would think that in the former case the vowels would be segregated by the independence of their changes and in the latter case they would be fused and be hard to hear individually. However, no difference in the perceived prominence of the individual vowels was found in the two conditions. This called into question the idea that the auditory system uses parallel changes to group components and independent changes to segregate them. Another negative finding was reported by Gardner and Darwin (1986). They gave one of the harmonics in a synthesized vowel a different pattern of frequency change from the others and found that it did not reduce the contribution of that harmonic to the identity of the vowel.

There would be a problem of interpretation even if we found that the parallel movement of partials promoted their fusion and that non-parallel movement promoted their segregation. Consider a sound that is harmonic (i.e. all its partials are multiples of its fundamental). Suppose all the partials are to be changed in frequency. When all of them are changed in parallel (on a logarithmic frequency scale), they remain multiples of a common fundamental whose frequency is changing. Therefore if the auditory system groups a subset of partials whenever they change in parallel, it may not be reacting to the parallel movement but to the continued presence of harmonic relations. Any experiment on the independent effect of parallel movement over and above the harmonicity factor has to be done with partials that are not harmonically related. To my knowledge, only one experiment of this type has shown a facilitation of the recognition of two separate sounds in a mixture. It was done by Chalikia and myself in an unpublished study. Vowels were synthesized using inharmonic spectra. Two were played at the same time and the listener

was asked to identify them. In some conditions they were more accurate when the partials of the two vowels glided in different paths than when they glided in parallel. However, the facilitating effect of independent motion was not large. Perhaps the auditory system's apparent neglect of this factor is due to the fact that frequency changes can be hard to detect in an environment in which there are reflections and reverberation. Hence they do not provide a robust clue to the structure of mixtures.

2.6 PHYSIOLOGICAL LIMITATIONS

In describing the use of these properties to group the components of the signal—properties such as the frequencies of components, the fit of all components into a harmonic series, their spatial origins, and so on—I have spoken as if there were no limits to the precision with which such properties and relations could be assessed by the auditory system. In actuality, this is not true. For example, assigning a separate estimate of spatial location may be difficult to achieve physiologically when the frequency components are closely packed or overlapped in the spectrum. Accordingly, the estimates would be unreliable. This is a good reason for using many clues to find the contributions of the individual acoustic sources to the incoming sound. Different ones may be reliably assessed under different acoustic circumstances. If they are allowed to compete and collaborate, the final decision should be more robust.

2.7 CONCLUSIONS

The principles of auditory scene analysis described in the foregoing paragraphs seem to resemble the principles of grouping described by the Gestalt psychologists. For example, these theorists described perceptual grouping in vision that was controlled by similarity in colour or proximity in space. Correspondingly, auditory grouping is promoted by similarity in timbre or in pitch, or by proximity in space. They also mentioned 'good continuation'. An example in vision is that when a contour is smooth, its parts are likely to be grouped and treated as the edge of a single object, whereas when a contour changes in a discontinuous way, its parts are less likely to be grouped. A corresponding example in audition is that if changes in the pitch of a voice are too sudden, as in the synthetic speech of Darwin and Bethell-Fox (1977) that I described earlier, the parts are not assigned to the same voice. There are some principles of grouping that seem to apply to only one sense or the other. For example bilateral symmetry in vision tends to group contours, but it would be stretching an analogy to apply this to audition. Conversely,

harmonic relations unite auditory components, but there is no analogy in vision. However, it is reasonable to conclude that the principles of grouping that were discovered and named by the Gestalt psychologists exist in order to perform the role of scene analysis. They serve, on the whole, to group sensory evidence that has been derived from the same (or closely related) environmental objects and events. Whatever correspondences exist between the principles that affect vision and audition do so because similar problems in the grouping of evidence are found in the two sense modalities. I have discussed the issue at length elsewhere (Bregman 1990).

The rules that group sounds by their sequential similarities may not always correctly bind together those that have come from the same environmental source, but they are likely to do so. When there are many rules, each with a good chance of grouping the right parts, a correct solution is likely to emerge if they are all permitted to 'vote' for the groupings that they favour, and the grouping with the highest number of votes is selected. A correct solution is probable as long as the system is operating in the rich natural environment in which it evolved. When it is placed in a soundproof room and presented with sounds in which many of the natural properties of the sound are missing, it is likely to produce strange illusions, such as the phenomena of auditory streaming and homophonic continuity that were discussed earlier, or the musical illusions described by Deutsch (e.g. 1975). These percepts represent the best the system can do when its strategies are evoked by unnatural data. Fortunately, these illusions display the nature of the rules to the experimental psychologist.

When I call the processes 'strategies' and describe them as solving problems, voting, and so on, I do not mean to imply that the components of the system know that they are doing these things. The metaphorical language that I have selected emphasizes the *contribution* that the various processes play in the adaptation of humans to their environments. Rather than linking the processes downward to a causal, physiological account of how they work, it links them upward, describing their functional role in the larger process of perceiving. Obviously both sorts of explanation, physiological and functional, are necessary for a full understanding.

I believe that there are a large number of strategies for grouping and interpreting the sensory data. The necessity for a large number comes from the fact that each one is subject to error. The one that tries to group sounds by their spatial origins may not be effective in reverberant environments. The one that groups partials only when they are harmonically related will fail when the event gives rise to inharmonic partials or to noisy sound that has no definite partials. The ones that look for sequential resemblances will sometimes be fooled by acoustic events that give rise to discontinuities in the sound. Because the clues are redundant,

all being the result of the same real-world events, the use of a number of relations will usually act as a protection against a failure of some of them. In the worst cases, listeners may simply not be able to penetrate the mixture of sounds. It is a testament to the power of the scene-analysis process that such total failures are quite rare.

REFERENCES

Bregman, A. S. (1978). Auditory streaming is cumulative. *Journal of Experimental Psychology: Human Perception and Performance*, **4**, 380–7.

Bregman, A. S. (1990). *Auditory scene analysis: the perceptual organization of sound*. MIT, Cambridge, MA.

Bregman, A. S. (1991). Using quick glimpses to decompose mixtures. In *Music, language, speech, and brain* (ed. J. Sundberg, L. Nord, and R. Carlson), pp. 284–93. MacMillan, London.

Bregman, A. S. and Campbell, J. (1971). Primary auditory stream segregation and perception of order in rapid sequences of tones. *Journal of Experimental Psychology*, **89**, 244–9.

Bregman, A. S. and Pinker, S. (1978). Auditory streaming and the building of timbre. *Canadian Journal of Psychology*, **32**, 19–31.

Bregman, A. S., Abramson, J., Doehring, P., and Darwin, C. J. (1985). Spectral integration based on common amplitude modulation. *Perception and Psychophysics*, **37**, 483–93.

Bregman, A. S., Liao, C., and Levitan, R. (1990*a*). Auditory grouping based on fundamental frequency and formant peak frequency. *Canadian Journal of Psychology*, **44**, 400–13.

Bregman, A. S., Levitan, R., and Liao, C. (1990*b*). Fusion of auditory components: effects of the frequency of amplitude modulation. *Perception and Psychophysics*, **47**, 68–73.

Brokx, J. P. L. and Noteboom, S. G. (1982). Intonation and the perceptual separation of simultaneous voices. *Journal of Phonetics*, **10**, 23–36.

Ciocca, V. and Bregman, A. S. (1989). Effects of auditory streaming on duplex perception. *Perception and Psychophysics*, **46**, 39–48.

Cutting, J. E. (1976). Auditory and linguistic processes in speech perception: Inferences from six fusions in dichotic listening. *Psychological Review*, **83**, 114–40.

Dannenbring, G. L. and Bregman, A. S. (1976). Stream segregation and the illusion of overlap. *Journal of Experimental Psychology: Human Perception and Performance*, **2**, 544–55.

Darwin, C. J. (1981). Perceptual grouping of speech components differing in fundamental frequency and onset-time. *Quarterly Journal of Experimental Psychology*, **33A**, 185–207.

Darwin, C. J. (1984). Perceiving vowels in the presence of another sound: constraints on formant perception. *Journal of the Acoustical Society of America*, **76**, 1636–47.

Darwin, C. J. and Bethell-Fox, C. E. (1977). Pitch continuity and speech source attribution. *Journal of Experimental Psychology: Human Perception and Performance*, **3**, 665–72.

Darwin, C. J. and Gardner, R. B. (1986). Mistuning a harmonic of a vowel: grouping and phase effects on vowel quality. *Journal of the Acoustical Society of America*, **79**, 838–45.

Deutsch, D. (1975). Musical illusions. *Scientific American*, **233**, 92–104.

Divenyi, P. L. and Oliver, S. K. (1989). Resolution of steady-state sounds in simulated auditory space. *Journal of the Acoustical Society of America*, **85**, 2042–52.

Dowling, W. J. (1973). The perception of interleaved melodies. *Cognitive Psychology*, **5**, 322–7.

Gardner, R. B. and Darwin, C. J. (1986). Grouping of vowel harmonics by frequency modulation: absence of effects on phonemic categorisation. *Perception and Psychophysics*, **40**, 183–7.

Guzman, A. (1969). Decomposition of a visual scene into three-dimensional bodies. In *Automatic interpretation and classification of images* (ed. A. Grasselli), pp. 243–276. Academic, New York.

Hall, J. W., Haggard, M. P., and Fernandes, M. A. (1984). Detection in noise by spectro-temporal pattern analysis. *Journal of the Acoustical Society of America*, **76**, 50–6.

Hartmann, W. M. and Johnson, D. (1991). Stream segregation and peripheral channeling. *Music Perception*, **9**, 155–84.

McAdams, S. (1984). *Spectral fusion, spectral parsing, and the formation of auditory images*. Ph.D. thesis, Stanford University. Stanford, CA.

McAdams, S. (1989). Segregation of concurrent sounds. I: Effects of frequency modulation coherence. *Journal of the Acoustical Society of America*, **86**, 2148–59.

Moore, B. C. J. (1989). *An introduction to the psychology of hearing* (3rd edn). Academic, London.

Moore, B. C. J. (1990). Co-modulation masking release: Spectro-temporal pattern analysis in hearing. *British Journal of Audiology*, **24**, 131–7.

Moore, B. C. J. and Patterson, R. D. (ed.) (1986) *Auditory frequency selectivity*, NATO-ASI Series. Plenum, New York.

Moore, B. C. J., Glasberg, B. R., and Peters, R. W. (1986). Thresholds for hearing mistuned partials as separate tones in harmonic complexes. *Journal of the Acoustical Society of America*, **80**, 479–83.

Scheffers, M. T. M. (1983). *Sifting vowels: auditory pitch analysis and sound segregation*. Ph.D. thesis, Groningen University. Groningen, The Netherlands.

Schubert, E. D. and Schultz, M. C. (1962). Some aspects of binaural signal selection. *Journal of the Acoustical Society of America*, **34**, 844–9.

Shepard, R. N. (1981). Psychophysical complementarity. In *Perceptual organization* (ed. M. Kubovy and J.R. Pomerantz), pp. 279–341. Erlbaum, Hillsdale, NJ.

Singh, P. (1987). Perceptual organization of complex-tone sequences: a tradeoff between pitch and timbre? *Journal of the Acoustical Society of America*, **82**, 886–99.

Smith, J., Hausfeld, S., Power, R. P., and Gorta, A. (1982). Ambiguous musical figures and auditory streaming. *Perception and Psychophysics*, **32**, 454–64.

Steiger, H. and Bregman A. S. (1981). Capturing frequency components of glided tones: frequency separation, orientation and alignment. *Perception and Psychophysics*, **30**, 425–35.

van Noorden, L. P. A. S. (1975). *Temporal coherence in the perception of tone sequences.* Ph.D. thesis, Eindhoven University of Technology. Eindhoven, The Netherlands.

van Noorden, L. P. A. S. (1982). Two channel pitch perception. In *Music, mind and brain: the neuropsychology of music* (ed. M. Clynes), pp. 251–70. Plenum, New York.

Warren, R. M. (1982). *Auditory perception: a new synthesis.* Pergamon, New York.

Warren, R. M. (1984). Perceptual restoration of obliterated sounds. *Psychological Bulletin*, **96**, 371–83.

Warren, R. M., Obusek, C. J., and Ackroff, J. M. (1972). Auditory induction: perceptual synthesis of absent sounds. *Science*, **176**, 1149–51.

3

Perception of acoustic sequences: global integration versus temporal resolution

Richard M. Warren

3.0 INTRODUCTION

Perception of acoustic sequences has long been a topic of major interest in psychoacoustics, due no doubt to the fact that speech and music consist of a succession of particular sounds occurring in specific orders. While the number of sounds employed in speech and the number employed in music are limited, their arrangements into extended sequential patterns are limitless. It seems quite reasonable to assume that comprehension of speech and appreciation of music could not be accomplished without the ability to identify component sounds and their order at some level of analysis. Thus, Hirsh (1959) related his measurements of thresholds for identification of order for pairs of items (consisting of sounds such as hisses, tones, and clicks) to the requirements for temporal resolution in speech perception. To illustrate these requirements, he cited the need for listeners to identify the order of /s/ and /t/ in order to discriminate between the words 'mist' and 'mitts'. More recently, Miller and Dexter (1988) stated: 'A major goal of research on speech perception is to explain how a listener derives the phonetic structure of an utterance during the course of language processing'. Similar statements have been made for music, with Winckel (1967) stating that when notes in music were played too rapidly, a perceptual metathesis (or permutation) of the order of the notes occurs, so that melody recognition becomes impossible.

However, there is a mounting body of evidence indicating that the comprehension of speech and the appreciation of music does not require their resolution into an ordered sequence of components, but rather involves global or holistic organization. It appears that this ability to perceive complex acoustic patterns globally occurs not only with verbal and musical sequences, but with sequences of arbitrarily selected brief

Thinking in sound: the cognitive psychology of human audition, ed. S. McAdams and E. Bigand. Oxford University Press, 1993, pp. 37–68.