8

Hearing Over Time

Information in the auditory domain is carried mainly by *changes* in the characteristics of sounds over time. This is true on a small time scale, when interpreting individual speech sounds; and on a larger time scale, when hearing the engine of a car become gradually louder as it approaches. However, it is the speed at which the auditory system can process sounds that is really remarkable. In free-flowing speech, consonants and vowels may be produced at rates of thirty per second (see Section 11.2.1). In order to process such fast-changing stimuli the auditory system has to have good *temporal resolution*.

This chapter examines two aspects of hearing over time. First, our ability to follow rapid changes in a sound over time, and second, our ability to combine information about sounds over much longer durations to improve detection and discrimination performance.

8.1 TEMPORAL RESOLUTION

Temporal resolution or temporal acuity refers to the resolution or separation of events in time. Although our ability to extract the frequency of a pure tone implies that the auditory system has some representation of the fine structure of a sound, we

do not hear the individual pressure variations as fluctuations in magnitude. When a continuous pure tone is played, we hear a continuous stable sound. Our perception of the magnitude of a sound is linked to the *envelope* (see Section 2.5.1), the variations in the peak pressure of a sound as a function of time. Temporal resolution usually refers to our ability to respond to rapid fluctuations in the envelope. If the response is sluggish, then the auditory system will not be able to track rapid fluctuations. The internal representation of the sound will be blurred in time, just as an out-of-focus visual image is blurred in space (and just as the excitation pattern is a blurred representation of the physical spectrum).

8.1.1 Measures of Temporal Resolution

Most of the early experiments on temporal resolution tried to measure a single duration that described the briefest change in a stimulus that can be perceived. The design of these experiments was complicated by the fact that any change in the *temporal* characteristics of a stimulus automatically produces changes in the *spectral* characteristics of the stimulus. We encountered this in Section 2.3.2 as spectral splatter. For example, one of the most popular temporal resolution experiments is the gap detection experiment. In this experiment, the listener is required to discriminate between a stimulus that is uninterrupted, and a stimulus that is the same in all respects except that it contains a brief silent interval or gap, usually in the temporal center of the stimulus (see Fig. 8.1). By varying the gap duration, it is possible to find the smallest detectable gap (the gap threshold). Unfortunately, introducing a sudden gap in a pure tone or other narrowband stimulus will cause a spread of energy to lower and higher frequencies. Hence, the gap may be detected by the auditory system as a change in the spectrum, rather than as a temporal event per se. Leshowitz (1971) showed that the minimum detectable gap between two clicks is only 6 *micro*seconds. However, this task was almost certainly performed using differences in the spectral energy at high frequencies associated with the introduction of the gap.

To avoid this confound, some researchers have measured gap detection for white noise (Fig. 2.15), the spectrum of which is not affected by an abrupt discontinuity.

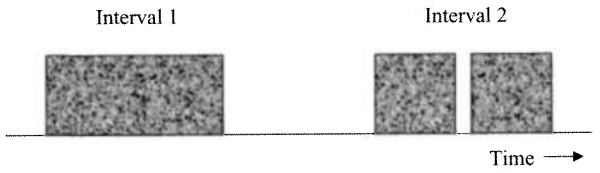

FIG. 8.1. The stimuli for a typical gap detection experiment. The listener's task is to pick the observation interval that contains the sound with the temporal gap (interval 2, in this case). The location of the gap (interval 1 or interval 2) would be randomized from trial to trial.

The gap threshold for white noise is around 3 ms (Penner, 1977). Another way to avoid spectral cues is to *mask* the spectral splatter with noise. Shailer and Moore (1987) measured gap detection for pure tones presented in a band-stop noise to mask the spectral splatter. They measured thresholds of about 4–5 ms that were roughly independent of the frequency of the tone, at least for frequencies above 400 Hz.

The magnitude spectrum of a sound is the same whether a sound is played forward or backward. Taking advantage of this property, Ronken (1970) presented listeners with two pairs of clicks. In one pair, the first click was higher in amplitude than the second, and in the other pair the second click was higher in amplitude than the first. The stimuli were therefore mirror images of each other in the time domain. Ronken found that listeners could discriminate between these stimuli when the gap between the clicks was just 2 ms. Taken together with the gap detection experiments, it appears that we can detect a change in level lasting only about 2–5 ms. As we discover in Section 8.2.1, our sensitivity to *repetitive* envelope fluctuations in sounds is even greater than that suggested by the results in this section.

8.1.2 Forward and Backward Masking

We have seen that when two sounds of similar frequency are presented together, the more intense sound may mask or obscure the less intense sound so that the less intense sound is inaudible. The masking effect extends *over time*, so that masking may be caused by a masker presented just before (*forward masking*) or just after (*backward masking*) the signal. Forward and backward masking are sometimes called *non-simultaneous* masking, because the masker and the signal do not overlap in time. Backward masking is a weak effect in trained listeners, and only causes an increase in the lowest detectable level of the signal when the signal is within 20 ms or so of the onset of the masker (Oxenham & Moore, 1994). Forward masking, however, can persist for over 100 ms after the offset of the masker (Jesteadt, Bacon, & Lehman, 1982). These effects can be regarded as aspects of temporal resolution, because they reflect our limited ability to "hear out" sounds presented at different times.

Figure 8.2 shows the smallest detectable level of a signal presented after a forward masker. The data are plotted as a function of the masker level, and as a function of the gap or silent interval between the masker and the signal. As the masker level is increased, the level of the signal required also increases. As the gap is increased, the masking decays, so that lower-level signals can be detected. Looking at the left panel, note that for low signal levels (below about 30 dB SPL), the growth of masking is shallow, so that a large change in masker level produces only a small change in the level of the signal at threshold. At higher levels, the growth of masking is roughly linear (1:1). These effects are reflected in the decay of forward masking with time. At high levels, the decay is faster than at low levels (right panel).

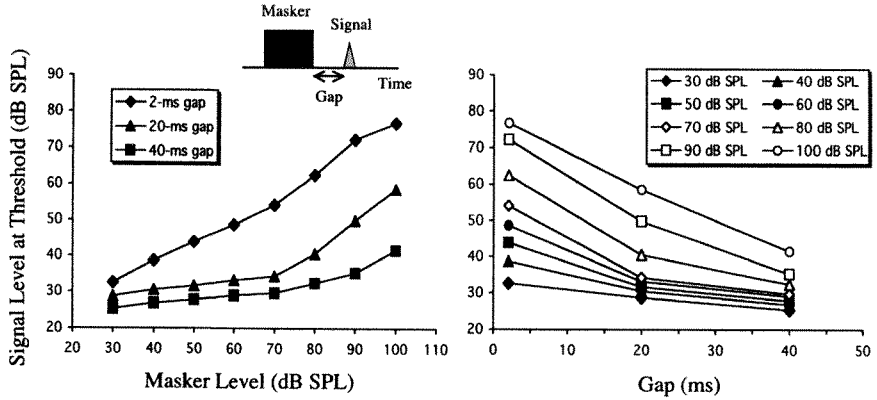

FIG. 8.2. The just-detectable level of a signal presented after a masker, as a function of the masker level and as a function of the gap between the masker and the signal. The masker and the signal were both 6000-Hz pure tones. Data are from Plack and Oxenham (1998).

It is now believed that forward and backward masking may be strongly influenced by the non-linearities in the cochlea discussed in Section 5.2.3. Recall that the response of the basilar membrane to a tone at characteristic frequency is roughly linear at low levels but compressive (shallow growth) at higher levels. For the conditions in Fig. 8.2, the masker falls within the compressive region of the response function: A 10-dB increase in physical masker level may produce only a 2-dB increase in the vibration of the basilar membrane. If the signal level is below about 30 dB SPL, it falls within the steeper, linear region of the response function, so a 2-dB increase in the vibration of the basilar membrane may require only a 2-dB increase in the physical signal level. It follows that the signal level needs to be increased by much less than the masker level to remain detectable, and the masking function (left panel) has a shallow slope. When the signal level is above 30 dB SPL, both the signal and the masker are compressed, so that the effects cancel out, and the result is linear growth in the masking function. The apparent rapid decay of masking at high signal levels is a result of the same mechanism (right panel). Suppose that a given increase in gap results in a constant reduction in the level of basilar-membrane vibration required for the signal to be detected. When the signal is at high levels, in the compressive region of the response, a given reduction in basilar membrane vibration will be associated with a much larger reduction in the physical signal level. The result is a steep reduction in signal threshold with time.

The response of the basilar membrane to a tone well below characteristic frequency is roughly linear. One might expect, therefore, the growth of forward masking with masker level to look very different when the masker frequency is below the signal frequency, and, indeed, it does. When the masker is below the

signal frequency and the signal level is within the compressive region, the growth of masking is very steep (an example of the *upward spread of masking*, see Section 5.4.4). A given change in masker level requires a much larger change in signal level, because the signal is compressed and the masker is not. Based on this reasoning, forward masking is currently being used to estimate the response of the human basilar membrane (Nelson, Schroder, & Wojtczak, 2001; Oxenham & Plack, 1997).

8.1.3 What Limits Temporal Resolution?

The auditory system is very good at representing the temporal characteristics of sounds. But, it is not perfect. What aspect of auditory processing limits temporal resolution, and is there a reason why resolution should be limited?

The temporal response of the auditory filter (see Section 5.2.2) is a potential limitation on temporal resolution. The basilar membrane continues to vibrate for a few milliseconds after the stimulus has ceased, effectively extending the stimulus representation in time, and smoothing temporal features such as gaps. Because the auditory filters are narrower at low center frequencies, the temporal response is longer at low frequencies than at high (see Fig. 5.3). If temporal resolution were limited by filter ringing, then we would expect resolution to be much worse at low frequencies than at high frequencies. However, the gap detection threshold for pure tones is roughly constant as a function of frequency, except, perhaps, at very low frequencies. Similarly, the decay of forward masking does not vary greatly with frequency. We do not have the hyper-acuity at high frequencies that may be suggested by the brief impulse response at 4000 Hz in Fig. 5.3. It is unlikely, therefore, that the temporal response of the basilar membrane contributes to the resolution limitation for most frequencies.

It has been suggested that forward masking is a consequence of neural adaptation. In Section 4.4.2, I describe how the firing rate in an auditory nerve fiber decreases with time after the onset of a sound. After the sound is turned off, the spontaneous firing rate is reduced below normal levels for 100 ms or so. The fiber is also less sensitive during this period of adaptation. The firing rate in response to a second sound will be *reduced* if it is presented during this time period. If adaptation is strong enough to push the representation of the second sound below its effective absolute threshold, then this could provide an explanation for forward masking. Although adaptation in the auditory nerve does not seem to be strong enough to account for the psychophysical thresholds (Relkin & Turner, 1988), adaptation at some other stage in the auditory system may be sufficient. Indeed, some physiologists make the implicit assumption that this is the case, and regard "adaptation" and "forward masking" as synonymous (much to my annoyance!).

Even if forward masking is partly a consequence of neural adaptation, adaptation cannot account for backward masking, in which the signal precedes the masker, and it cannot readily account for gap detection: An overall reduction in firing rate may have little effect on the internal representation of a 5-ms gap. The temporal resolution limitation may be better explained by an *integration* mechanism which combines or sums neural activity over a certain period of time. Such a mechanism would necessarily produce a limitation in temporal resolution, because rapid fluctuations would effectively be "averaged out" as they are combined over time. The integration mechanism would also result in a persistence of neural activity, because after the stimulus had been turned off, the mechanism would still be responding to activity that had occurred earlier. An analogy is the electric hob on a cooker. The hob takes time to warm up after the electricity has been switched on, and time to cool down after the electricity has been switched off. The temperature at a given time is dependent on a weighted integration of the electric power that has been delivered previously.

Some neurons in the auditory cortex have sluggish responses that may provide the neural substrate for such an integration device. In general, however, the integration time may arise from the processing of information in the central auditory system. Neural spikes are "all or nothing" in that each spike has the same magnitude. Information is carried, not by the *magnitude* of each spike, but by the number of spikes per second, or by the temporal regularity of firing in the case of phase-locking information. To measure intensity or periodicity, it is necessary to *combine* information over time: To count the number of spikes in a given time period, or, perhaps, to compute an autocorrelation function over a number of delays. Both these processes imply an integration time that will necessarily limit temporal resolution. A model of the integration process is described in Section 8.1.4.

8.1.4 The Temporal Window

The temporal window model is a model of temporal resolution that is designed to accommodate the results from most temporal resolution experiments, although the model parameters are based on forward and backward masking results. The stages of the model are shown in Fig. 8.3. The first stage of the model is a simulation of the auditory filter, which includes the non-linear properties discussed in Chapter 5. This stage provides a simulation of the vibration of a single place on the basilar membrane. The second stage is a device that simply squares the simulated velocity of vibration. Squaring has the benefit of making all the values positive, but it may also reflect processes in the auditory pathway. Finally, the representation of the stimulus is smoothed by the *temporal window*, a sliding temporal integrator.

The temporal window is a function that *weights* and *sums* the square of basilar membrane velocity over a short time period, and is assumed to reflect processes in the central auditory system. The temporal window has a center time, and times close to the center of the window receive more weight than times remote from the center of the window, just like the auditory filter in the frequency domain. (Indeed, there is a temporal equivalent of the ERB called the *equivalent rectangular*

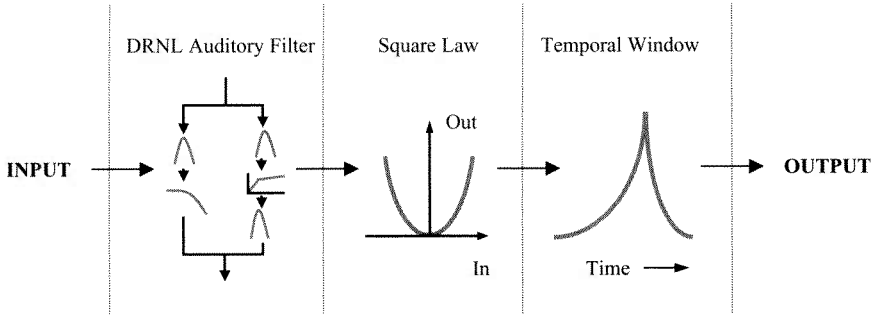

FIG. 8.3. The latest version of the temporal window model, comprising a simulation of the auditory filter (i.e., the response of a single place on the basilar membrane), a device that squares the output of the filter, and the temporal window itself, which smooths the output of the cochlear simulation. The output of the model represents the internal representation of the input for a single frequency channel. The figure is redrawn from Plack, Oxenham, and Drga (2002). The auditory filter design is based on Meddis, O'Mard, and Lopez-Poveda (2001).

duration, ERD, of the window. The ERD is around 8 ms and is assumed not to vary with frequency, because measures of temporal resolution such as gap detection do not vary with frequency.) Times before the center of the window are given more weight than times after the center of the window, to reflect the greater effectiveness of forward compared to backward masking. Thus, at any instant, the output of the temporal window is a *weighted average* or *integration* of the intensity of basilar-membrane vibration for times before and after the center of the window. A smoothed representation of the stimulus is derived by calculating the output of the temporal window as a function of center time. This is called the *temporal excitation pattern* (TEP), and is analogous to the excitation pattern in the frequency domain. The TEP is a description of how variations in level are represented in the central auditory system.

Figure 8.4 shows the output of the temporal window model for a signal preceded by a forward masker. The shallow skirts of the temporal window for times before the center mean that the abrupt offset of the physical masker is represented internally as a shallow decay of excitation. The model suggests that the neural activity produced by the masker *persists* after the masker has been turned off. If the masker has not decayed fully by the time the signal is presented, then, effectively, the signal is masked *simultaneously* by the residual masker excitation. It is assumed that the listener only has access to the TEP produced by the signal and the masker combined, and so the signal may be detected by virtue of the bump on the combined TEP.

The single-value measures of temporal resolution described in Section 8.1.1 may not appear at first glance to be consistent with the much longer time scale

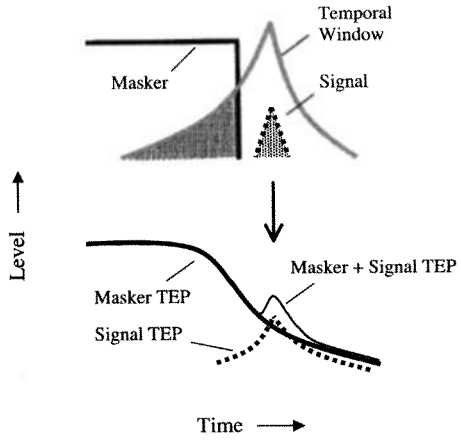

FIG. 8.4. An illustration of how the temporal window model accounts for forward masking. A temporal window centered on the time of occurrence of a brief signal integrates most of the signal and some of the forward masker (top panel). The integrated masker energy acts to mask the signal just as if the masker and signal were simultaneous. The bottom panel shows temporal excitation patterns (TEPs) for the signal, the masker, and the TEPs for the masker and signal added together. The TEP is the output of the temporal window as a function of the center time of the window, and is a "smoothed" version of the original waveform. The output of the temporal window in the top panel represents a single instant on the TEP (indicated by the arrow).

associated with the interactions between the masker and the signal in forward masking. However, the temporal window model shows us that there are two components limiting performance on the standard temporal resolution task. The first is the degree to which the representation of the stimulus is smoothed by the sluggish response of the system, as modeled by the temporal window. The second is our ability to detect the fluctuation in level in the smoothed representation (or at the output of the temporal window). These tasks involve an *intensity discrimination* component as well as the "pure" temporal resolution component. Because the signal level at threshold in a forward masking task is usually much less than the masker level, it has the effect of producing only a small bump in the otherwise smooth decay of masker excitation, even though the gap between the masker and the signal may be several tens of milliseconds. On the other hand, in a typical gap detection experiment, the change of level that produces the gap is very great, and so it can be detected at shorter time intervals. The temporal window model predicts that the bump in the TEP corresponding to the signal in forward masking (see Fig. 8.4), and the dip in the TEP corresponding to the gap in the gap detection task (see Fig. 8.5) should be roughly the same size at threshold.

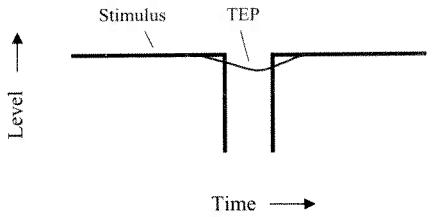

FIG. 8.5. The output of the temporal window model in response to a sound containing a temporal gap. A temporal window centered on the gap integrates stimulus energy from before and after the gap, so that the period of absolute silence in the original stimulus is replaced by a shallow dip in excitation level in the internal representation.

Because the first stage in the simulation is an auditory filter with a center frequency that can be allowed to vary, the temporal window model illustrated in Fig. 8.3 represents the output of just a single frequency channel. However, if the TEP is calculated for a number of center frequencies, a description of the internal representation of a stimulus across both frequency and time can be obtained. This is a three-dimension plot called a *spectro-temporal excitation pattern* (STEP). Fig. 8.6 shows the STEP for the utterance "bat." A cross-section across frequency at a given time provides an excitation pattern, and shows the blurring in the frequency domain produced by the auditory filters. A cross-section across time at a given frequency provides a temporal excitation pattern, and shows the blurring in the time domain produced by the temporal window. The STEP is, therefore, an estimate of the resolution of the auditory system with respect to variations in level across both frequency and time. As a rule of thumb, if a stimulus feature can be seen in the STEP, then we can probably hear it—if not, then we probably can not.

8.1.5 Across-Channel Temporal Resolution

It is important that we are able to track rapid changes occurring at a particular frequency. It is also important that we are sensitive to the timing of events *across* frequency. We need the latter ability to detect frequency sweeps such as formant transitions in speech (see Section 11.1.1), and to segregate sounds on the basis of differing onset or offset times (see Section 10.2.1). Pisoni (1977) reported that we are capable of detecting a difference in the onset times of a 500-Hz pure tone and a 1500-Hz pure tone of just 20 ms. With evidence of even greater resolution, Green (1973) reported that listeners could discriminate delays in a limited spectral region of a wideband click down to 2 ms. On the other hand, our ability to detect a temporal gap between two pure tones separated in frequency is very poor, with gap thresholds of the order of 100 ms (Formby & Forrest, 1991). In this latter case, however, listeners are required to judge the time interval between an offset and an

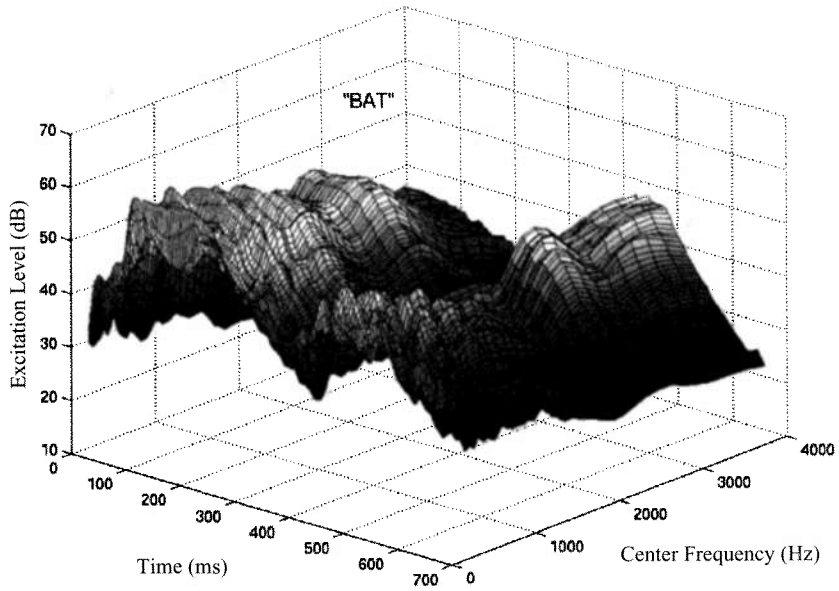

FIG. 8.6. The spectro-temporal excitation pattern for the utterance "bat." The low-frequency resolved harmonics and the upper formant peaks (composed of unresolved harmonics) can be seen for times up to around 300 ms after the start. The "t" sound produces excitation at high center frequencies around 500 ms after the start.

onset (rather than between two onsets). It could be that the auditory system has no particular ecological reason to be good at this, and, hence, has not evolved or developed the appropriate neural connections.

A difficulty with some of these experiments is avoiding "within-channel" cues. For example, in an across-frequency gap detection experiment, the output of an auditory filter with a center frequency *between* the two tones may show a response to both tones so that the representation is not dissimilar to the standard gap-detection task in which both tones have the same frequency. Although it is perfectly valid for the auditory system to detect across-frequency temporal features by the activity in a single frequency channel, it does mean that the experiments may not be measuring across-*channel* (or across-characteristic-frequency) processes in every case.

8.2 THE PERCEPTION OF MODULATION

Our ability to detect a sequence of rapid fluctuations in the envelope of a stimulus can also be regarded as a measure of temporal resolution. However, I have given modulation a separate section, because there are some phenomena associated with

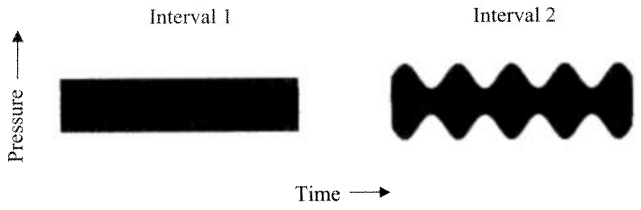

FIG. 8.7. The stimuli for a typical modulation detection experiment. The listener's task is to pick the observation interval that contains the sound with the modulated envelope (interval 2, in this case). The location of the modulation (interval 1 or interval 2) would be randomized from trial to trial.

modulation perception that go beyond the question of how rapidly the system can react to changing stimuli.

8.2.1 The Modulation Transfer Function

In a typical modulation detection task, the listener is required to discriminate a noise or tone that is sinusoidally amplitude modulated (see Section 2.5.1) from one that has a flat envelope (see Fig. 8.7). A plot of the smallest detectable depth of modulation against the *frequency* of modulation describes a *modulation transfer function* for the auditory system. The choice of carrier (the sound that is being modulated) is very important. If a pure-tone carrier is used, then care must be taken to ensure that the spectral side bands—the two frequency components either side of the carrier frequency—are not resolved on the basilar membrane. If they are resolved, then listeners may perform the modulation detection task using features in the excitation pattern, rather than by a temporal analysis of the envelope. Because the frequency difference between the carrier and each side band is equal to the modulation rate, the use of pure tone carriers is limited to relatively low modulation rates. The highest modulation frequency that can be used depends on the carrier frequency, because the bandwidth of the auditory filter increases (and, hence, the resolving power of the basilar membrane decreases) as center frequency is increased. Thus higher modulation frequencies can be used with higher frequency carriers.

When the carrier is a white noise, there are no long-term spectral cues to the presence of the modulation. In this case, listeners show roughly equal sensitivity to amplitude modulation for frequencies of up to about 50 Hz, and then sensitivity falls off (see Fig. 8.8). The modulation transfer function has a *low-pass* characteristic, and behaves like a low-pass filter in the envelope or modulation domain. As expected from the temporal resolution results in Section 8.1.1, the auditory system cannot follow fluctuations that are too fast. However, when the modulation depth is 100% (i.e., the envelope goes right down to zero in the valleys) we are able to detect modulation frequencies as high as *1000 Hz*! The auditory system is

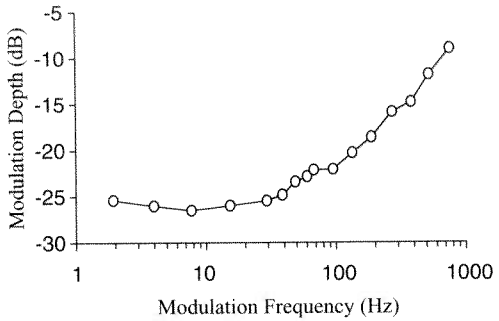

FIG. 8.8. A temporal modulation transfer function, showing the smallest detectable depth of sinusoidal amplitude modulation, imposed on a white noise carrier, as a function of the frequency of modulation. The lower the modulation depth at threshold, the greater the sensitivity to that frequency of modulation. Modulation depth is expressed as 20 log (m), where m is the modulation index (see Section 2.5.1). On this scale, 0 dB represents 100% modulation (envelope amplitude falls to zero in the valleys). Data are from Bacon and Viemeister (1985).

much faster than the visual system in this respect. The maximum rate of flicker detectable by the visual system is only about 50 Hz. The highest detectable modulation frequency for a sound (corresponding to a period of only 1 ms) is higher than would be expected from the gap-detection data, and suggests that the auditory system is more sensitive to *repeated* envelope fluctuations (as in the modulation detection task) than to a single envelope fluctuation (as in the gap detection task).

A problem with using noise as a carrier is that it contains "inherent" envelope fluctuations: The envelope of noise fluctuates randomly in addition to any modulation added by the experimenter. These inherent fluctuations may obscure high-frequency modulation that is imposed on the carrier. If sinusoidal carriers, which have flat envelopes, are used instead, then the high-sensitivity portion of the modulation transfer function extends up to about 150 Hz, rather than just 50 Hz (Kohlrausch, Fassel, & Dau, 2000).

8.2.2 Modulation Interference and the Modulation Filterbank

The idea that modulation processing can be characterized by a low-pass filter in the envelope domain, with low modulation frequencies passed and high modulation frequencies attenuated, may be too simplistic. The implication is that all envelope fluctuations are processed together. However, our ability to hear one pattern of modulation in the presence of another pattern of modulation depends on the frequency separation of the different modulation frequencies. If they are far removed in modulation frequency, then the task is easy. If they are close in modulation

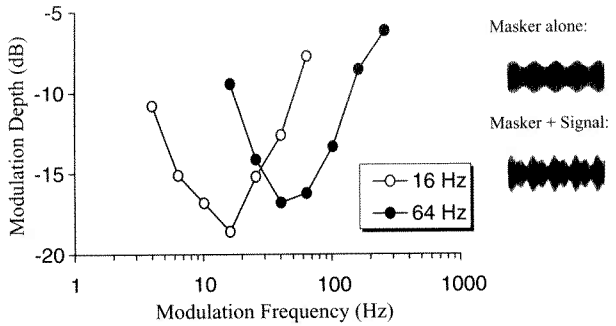

FIG. 8.9. Two psychophysical tuning curves in the envelope domain. The curves show the modulation depth of "masker" modulation required to mask "signal" modulation, as a function of the modulation frequency of the masker. The legend shows the modulation frequency of the signal. Both masker and signal modulation were imposed on the *same* noise carrier (see schematic on the right). The modulation depth of the signal was fixed at −15 dB for the 16-Hz signal modulation, and at −17 dB for the 64-Hz signal modulation. The curves show tuning in the envelope domain. When the masker modulation frequency is remote from the signal modulation frequency, then the auditory system can separate the two modulation patterns, and a high modulation depth is required to mask the signal. Data are from Ewert and Dau (2000).

frequency, then the task is hard (Fig. 8.9). The auditory system exhibits *frequency selectivity* in the envelope domain, just as it does in the fine-structure domain. Indeed, the two types of analysis may be independent to some extent: Interference between nearby modulation frequencies occurs even if the *carrier* frequencies are very different (Yost, Sheft, & Opie, 1989).

Dau and colleagues (Dau, Kollmeier, & Kohlrausch, 1997) argue that the auditory system contains a bank of overlapping "modulation filters" (analogous to the auditory filters) each tuned to a different modulation frequency. Just as we can listen to the auditory filter centered on the signal frequency, thus attenuating maskers of different frequencies, we also may be able to listen to the *modulation* filter centered on the *modulation* frequency we are trying to detect, and masker modulation frequencies remote from this may be attenuated. It has been estimated that the bandwidth of the modulation filters is roughly equal to the center modulation frequency, so that a modulation filter tuned to 20-Hz modulation has a bandwidth of about 20 Hz (Ewert & Dau, 2000). There is evidence that some neurons in the inferior colliculus are sensitive to different modulation frequencies (Langner & Schreiner, 1988), and these neurons may be the physiological substrate for the modulation filterbank.

Although there is controversy over the main premise, the modulation filterbank can account for many aspects of modulation perception, including the interference between different modulation frequencies described above. Looking at the

broader picture, it seems plausible that the auditory system may decompose a complex sound in terms of modulation frequency, as a source of information for determining sound identity. Furthermore, the modulation filterbank may help the auditory system to separate out sounds originating from different sources, which often contain different rates of envelope fluctuations (see Section 10.2.1).

8.2.3 Comodulation Masking Release

When we are trying to detect a pure-tone signal in the presence of a *modulated* masker, it helps if there are additional frequency components in a different part of the spectrum that have the *same pattern of envelope fluctuations* as the masker (Hall, Haggard, & Fernandes, 1984). These additional components are said to be *comodulated* with the masker, and the reduction in threshold when they are added is called *comodulation masking release*. Many experiments use a modulated noise band or pure tone centered on the signal as the on-frequency masker. Additional "flanking" noise bands or pure tones, with frequencies removed from the masker but with coherent modulation, can then be added to produce the masking release (see Fig. 8.10). Although some of the performance improvements may be the result of interactions between the flankers and the masker at a single place on the basilar

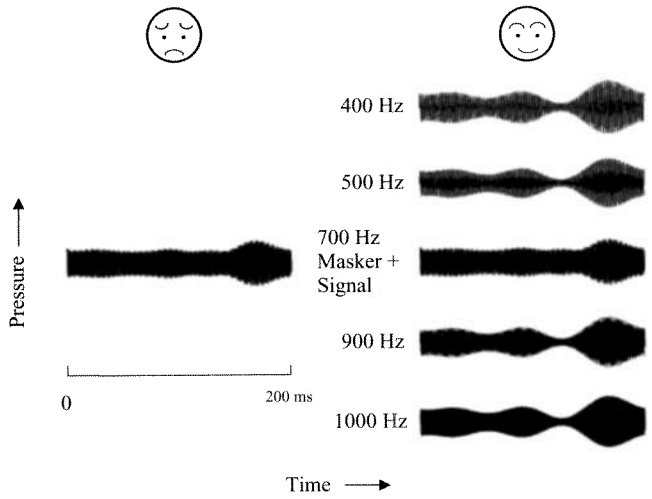

FIG. 8.10. Comodulation masking release. The signal is a 700-Hz pure tone, masked by a modulated 700-Hz pure tone. Signal detection is hard when the masker and the signal are presented on their own (left panel), but is easier with the addition of flanking tones with frequencies of 400, 500, 900, and 1000 Hz, comodulated with the masker (right panel). Note the distinct envelope of the 700-Hz band, caused by the presence of the signal.

membrane (*within-channel* cues), a comodulation masking release of around 7 dB can be produced when the flankers are far removed in frequency from the masker and the signal, and even when the flankers are presented in the *opposite ear* to the masker and the signal (Schooneveldt & Moore, 1987).

It appears that the auditory system has the ability to make comparisons of envelope fluctuations across frequency and across ears. When the signal is added to the masker, the pattern of envelope fluctuations will change slightly, and the signal may be detected by a disparity between the fluctuations produced by the masker and signal, and the fluctuations of the flanking bands (Richards, 1987). Alternatively, the auditory system may use a dip in the envelope of the flankers as a cue to the best time to listen for the signal (Buus, 1985): In the case of comodulated flankers, dips in the flanker envelope correspond to dips in the masker envelope, which is when the masker intensity is least. Like the modulation interference experiments, these experiments may be illustrating ways in which the auditory system uses envelope fluctuations to separate sounds from different sources. Sound components from a single source tend to have coherent envelope fluctuations across frequency, just like the comodulated flankers.

8.2.4 Frequency Modulation

Amplitude modulation and frequency modulation may seem to be very different aspects of dynamic stimuli. In the former, the amplitude is varying, and in the latter the frequency is varying. However, the distinction may not be so obvious to the auditory system. Consider the response of an auditory filter (or the response of a place on the basilar membrane) with a center frequency close to the frequency of a pure tone that is frequency modulated. As the frequency of the pure tone moves *toward* the center of the filter, the filter output will *increase*. As the frequency of the pure tone moves *away from* the center of the filter, the filter output will *decrease*. In other words, the output of the auditory filter will be *amplitude* modulated. Considering the whole excitation pattern, as the frequency moves up the excitation level will decrease on the low-frequency side, and increase on the high-frequency side, and conversely as the frequency moves down.

It is thought that for modulation frequencies of greater than about 10 Hz, frequency modulation and amplitude modulation are detected by the *same mechanism*, based on envelope fluctuations on the basilar membrane. At these rates, frequency modulation can interfere with the detection of amplitude modulation and vice versa (Moore, Glasberg, Gaunt, & Child, 1991). For lower modulation frequencies, the auditory system may be able to track the change in the pattern of phase locking associated with the variation in frequency. At low rates, therefore, detection of frequency modulation may be based more on *pitch* cues than on envelope cues. The fact that variations in pitch can be tracked only when they are fairly slow suggests that the pitch mechanism is quite sluggish, with relatively poor temporal resolution compared to temporal resolution for level changes.

8.3 COMBINING INFORMATION OVER TIME

In Section 6.3.5 I describe how the sensitivity of the auditory system to differences in intensity may be improved by combining the information from several nerve fibers. The same is true (at least theoretically) of combining information over *time*. In this section we explore how the auditory system may integrate information over time to improve performance on a number of auditory tasks.

8.3.1 Performance Improvements With Duration

For many of the mindless experiments that we pay people to endure, performance improves as the duration of the stimuli is increased. Figure 8.11 shows the effect of duration on our ability to detect a pure tone, to discriminate between the intensities of two pure tones, to discriminate between the frequencies of two pure tones, and to discriminate between the fundamental frequencies of two complex tones. In each task, performance improves rapidly over short durations, but, after a certain "critical" duration, there is little additional improvement as the stimulus duration is increased.

The range of durations over which performance improves differs between tasks. The critical duration can be as much as 2 seconds at 8000 Hz for intensity discrimination (Florentine, 1986). The duration effect depends on frequency for frequency discrimination: Low-frequency tones show a greater improvement with duration than high-frequency tones (Moore, 1973). Similarly, the duration effect increases as fundamental frequency decreases for fundamental frequency discrimination with unresolved harmonics (Plack & Carlyon, 1995). In addition, unresolved harmonics show a greater improvement with duration than do resolved harmonics, even when the fundamental frequency is the same (White & Plack, 1998). The following sections examine ways in which the auditory system may combine information over time in order to improve performance.

8.3.2 Multiple Looks

Imagine that you are given a big bucket of mud containing a large number of marbles, and you are asked to decide whether there are more blue marbles or more red marbles in the bucket. You are allowed a certain number of opportunities to select marbles from the mud at random, and you can select four marbles each time, although you have to replace them in the bucket after you have counted them. Let's say on your first try, you pick out three blue marbles and one red. On your second try, you pick out four blue marbles and no red. On your third try, you pick out two blue marbles and two red. On your fourth try, you pick out three blue marbles and

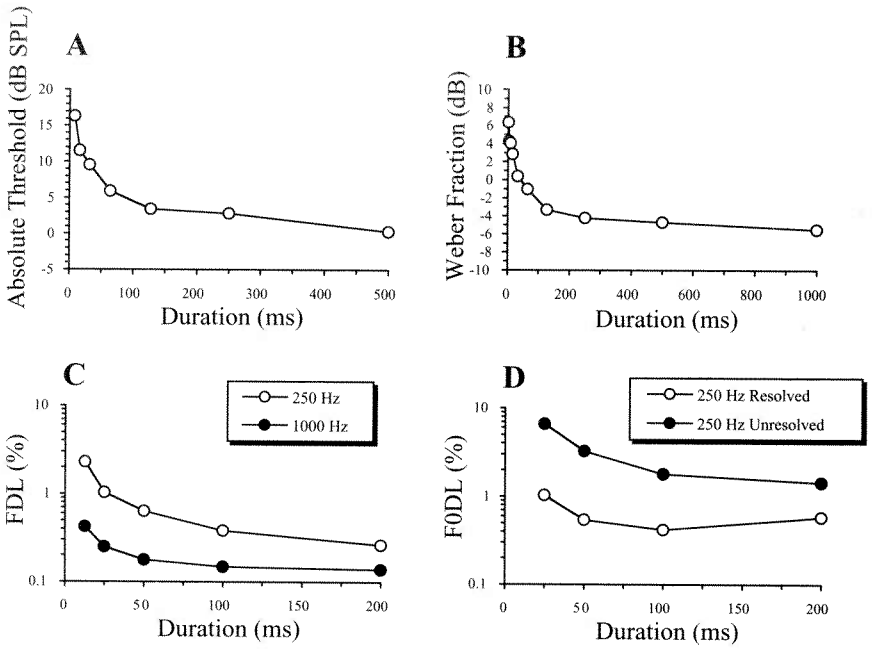

FIG. 8.11. Performance improvements with stimulus duration for four different auditory tasks: A) detection of a 1000-Hz pure tone (Florentine, Fastl, & Buus, 1988); B) intensity discrimination for a 1000-Hz pure tone (Florentine, 1986); C) frequency discrimination for 250- and 1000-Hz pure tones (Moore, 1973); and D) fundamental frequency discrimination for complex tones consisting of resolved or unresolved harmonics, both with a fundamental frequency of 250 Hz (Plack & Carlyon, 1995). FDL and F0DL refer to frequency difference limen and fundamental frequency difference limen respectively.

one red. Note that the more picks you have, the more confident you are that there are more blue marbles than red marbles. If you simply add up the total numbers of blue marbles and red marbles that you have removed, you can make your decision based on whichever color is most numerous. The more chances to pick you have, the greater is the accuracy of this final measure.

If you replace "decide whether there are more blue marbles or more red marbles" with "make a correct discrimination between sounds," you have the basis for the multiple looks idea. The more often you sample, or take a "look" at, a stimulus, the more likely you are to make a correct discrimination; to decide, for example, whether one sound is higher or lower in intensity than another sound. Discrimination performance is usually limited by the *variability* of our internal representation of the stimulus. The variability might be due to variations in the firing rates of neurons, or in the precision of phase locking, or it might be due

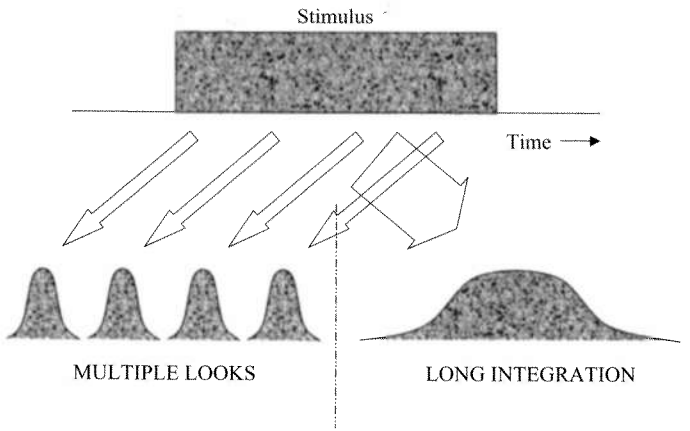

FIG. 8.12. A highly schematic illustration of how the information in a sound may be analyzed using several short-duration samples (left) or one long-duration sample (right).

to some variation in the stimulus itself (e.g., in the level of a noise). If we only sample a short time interval we might make a mistake because, just by chance, the number of neural spikes in the sample in response to one sound may be greater than the spikes in response to a slightly more intense sound. If we could add up the neural spikes from several samples, over a longer time interval, we would be much more likely to make an accurate judgment. Figure 8.12 shows how a long stimulus might be broken down into a number of discrete samples.

The multiple looks method will only work if the samples are *independent* of each other. It would be no use taking just one sample of marbles and multiplying the numbers of each color in that sample by ten. You would be just as likely to make a mistake as if you had stuck with the original numbers. Similarly, if you take two counts of the number of spikes over intervals that overlap in time, you will be counting the same spikes twice, and they won't help you any further. Sometimes, there is a problem even if the samples don't overlap. Imagine that you have a noise stimulus whose amplitude is modulated randomly—but slowly—up and down. If you take two samples close together in time, then they will not be independent. They might both fall on the same peak, for example. Your estimate of the *overall* level of the noise will not benefit greatly from the second sample, in this case.

The mathematics of combining information in this way were sorted out a while ago (see McNicol, 2004). The ability to make a discrimination between two stimuli can be expressed in terms of percent correct responses, but, also, in terms of the discrimination index, d′ ("d-prime"). d′ is a measure of a listener's ability to discriminate between two stimuli. d′ is equal to the difference between the internal representations of the stimuli, divided by the standard deviation of the representations. If two independent samples of the difference are simply added, then the

difference increases by a factor of two, but the standard deviation only increases by the square root of two. It follows that d′ increases by $\sqrt{2}$ (1.414). In general, d′ increases in proportion to the *square root* of the number of samples. Hence, if the duration of a stimulus is increased, which allows more independent samples, then performance should improve. Crucially, performance should not depend on *when* the samples are taken, so long as a memory representation of the earlier samples can be held until they are combined with the later ones.

Evidence for the use of multiple looks by the auditory system can be found in several tasks. Viemeister and Wakefield (1991) demonstrated that the detectability of a pair of short pure-tone bursts is greater than that for a single short tone burst, even when there is a noise between the bursts. Furthermore, performance is *independent* of the level of the noise. It appears that the auditory system can sample and combine the information from the two bursts, without including the intervening noise that would have acted as a powerful masker had it been integrated with the tones. White and Plack (1998) found that fundamental frequency discrimination performance for two 20-ms complex tone bursts separated by a brief gap (5–80 ms) is almost exactly that predicted by the multiple-looks hypothesis, when compared to performance for one tone burst. Performance is independent of the gap between the bursts, again consistent with the multiple-looks hypothesis. There is good evidence, therefore, that something similar to multiple-looks processing is used by the auditory system in some circumstances.

8.3.3 Long Integration

An alternative to taking several short-duration samples is to take one long-duration sample. Fig. 8.12 illustrates the distinction between multiple looks and long integration in the processing of a long stimulus. In some situations, it may be beneficial for the auditory system to perform an analysis on a long *continuous* chunk of the stimulus. Alternatively, it may be problematic for the auditory system to combine multiple samples from different times in an optimal way.

There is some evidence that the auditory system is able to obtain a benefit from continuous long integration that it would not get from combining a succession of discrete samples. The effect of increasing duration on performance is often much greater than would be predicted by the multiple-looks model, particularly over short durations (note the rapid improvements in performance over short durations in Fig. 8.11). For instance a doubling in the duration of an unresolved complex tone from 20 to 40 ms results in a threefold increase in d′ for fundamental frequency discrimination (White & Plack, 1998). This is much larger than the factor of $\sqrt{2}$ predicted by a doubling in the number of independent samples.

In Section 6.2.2, it is described how *loudness* increases with duration, for durations up to several hundred milliseconds. When we are judging the absolute magnitude of sounds, we seem to be able to combine level information over quite a long period, and this may involve some sort of long integration mechanism. Long integration may be useful to us when trying to determine, for example, whether a sound is getting closer or not. We need to be able to detect changes in the long-term magnitude of the sound, not in the rapid fluctuations in level that are characteristic of the particular sound that is being produced.

8.3.4 Flexible Integration

Viemeister and Wakefield (1991) reported that when two pure-tone bursts are separated by a gap of 5 ms or more, detection performance for the two tone bursts, compared to one tone burst, is consistent with multiple looks. However, when the bursts are less than five milliseconds apart, performance is improved further. Similarly, White and Plack (1998) found that fundamental frequency discrimination for a pair of complex-tone bursts is better when the tone bursts are continuous, rather than when there is a gap between the bursts. It is possible that the auditory system uses a long integration window for continuous stimuli, benefiting from the performance advantages, but resets the integration time when there is a discontinuity. In these situations, the two bursts may be integrated separately and the information combined (when necessary) using a multiple-looks mechanism.

It makes sense that the auditory system should be flexible in this way. A discontinuity is often indicative of the end of one sound feature and the beginning of another. These separate features may require separate analysis. It may not be optimal, for example, to average the pitches of two consecutive tones, when identification of the sound may depend on detecting a difference between them.

Furthermore, temporal resolution tasks, such as gap detection (or the detection of a stop consonant), require that a short integration time, possibly something similar to the temporal window, is used. An integration time of several hundred milliseconds would not be able to track a discontinuity of only 3 ms: Any brief dips or bumps would be smoothed out in the internal representation. *If* true long integration does exist, then shorter integration times must also exist. It is possible (and I am speculating freely here) that the different integration times may be implemented by auditory neurons with different temporal responses. Sluggish neurons could provide long integration, relatively fast neurons could provide short integration.

8.4 SUMMARY

As well as being able to detect very rapid changes in a stimulus, the auditory system is capable of combining information over much longer times to improve performance. The neural mechanisms underlying these abilities may represent a flexible response to the different temporal distributions of information in sounds.

1. Auditory temporal resolution is very acute. We can detect level changes lasting less than five milliseconds.

2. Forward and backward masking show that the influence of a stimulus is *extended over time*, affecting the detection of stimuli presented after or before. This influence may reflect a persistence of neural activity after stimulus offset (and a build-up in response after onset), which can be modeled using a sliding temporal integrator or *temporal window*. Alternatively, forward masking may be a consequence of the reduction in sensitivity associated with neural adaptation.

3. We are extremely sensitive to repetitive fluctuations. We can detect amplitude modulation at rates up to 1000 Hz.

4. One pattern of modulation may interfere with the detection of another pattern of modulation, but not when the modulation frequencies are very different. We may have specific neural channels that are tuned to different rates of modulation and behave like a *modulation filterbank*.

5. The addition of frequency components in different regions of the spectrum, but with the *same pattern of modulation* as a masker, can improve our ability to detect a signal at the masker frequency. This finding may reflect a general ability to use coherent patterns of modulation across frequency to separate out simultaneous sounds.

6. Frequency modulation may be detected by the induced amplitude modulation in the excitation pattern, for modulation frequencies above about 10 Hz. At lower rates, the frequency excursions may be tracked by a (sluggish) mechanism based on phase locking.

7. Performance improves in many hearing tasks as the duration of the stimulus is increased. These improvements may result from a multiple-looks mechanism that combines several short samples of the stimulus, or from a long integration mechanism, which analyzes a continuous chunk of the stimulus. Flexible integration times may allow the auditory system to respond to rapid changes in a stimulus, and to integrate over longer durations when necessary.

8.5 READING

The ideas expressed in this chapter are developed further in:

Viemeister, N. F., and Plack, C. J. (1993). Time analysis. In W. A. Yost, A. N. Popper & R. R. Fay (Eds.), *Human psychophysics* (pp. 116–154). New York: Springer-Verlag.

Eddins, D. A., and Green, D. M. (1995). Temporal integration and temporal resolution. In B. C. J. Moore (Ed.), *Hearing* (pp. 207–242). New York: Academic Press.

9

Spatial Hearing

"Where is that sound coming from?" is a question our brains often pose to our ears. Most sounds originate from a particular place because the *source* of most sounds is a vibrating object with a limited spatial extent. There are a number of reasons why we would like to be able to locate the source of a sound. First, the location of the sound may be important information in itself. For example, did the sound of distant gunfire come from in front or behind? Second, the location of the sound may be used to orient visual attention: If someone calls your name you can turn to see who it is. Finally, sound location can be used to separate out sequences of sounds arising from different locations, and to help us attend to the sequence of sounds originating from a particular location (see Chap. 10). Location cues can help us to "hear out" the person we are talking to in a room full of competing conversations. Similarly, location cues help us to hear out different instruments in an orchestra or in a stereo musical recording, adding to the clarity of the performance.

I must admit that the visual system is about 100 times more sensitive to differences in source location than is the auditory system. However, our eyes are limited in their field of view, whereas our ears are sensitive to sounds from any direction and from sound sources that may be hidden behind other objects. Much information about approaching danger comes from sound, not light. This chapter describes