13. Psychoacoustics

Psychoacoustics is concerned with the relationships between the physical characteristics of sounds and their perceptual attributes. This chapter describes: the absolute sensitivity of the auditory system for detecting weak sounds and how that sensitivity varies with frequency; the frequency selectivity of the auditory system (the ability to resolve or *hear out* the sinusoidal components in a complex sound) and its characterization in terms of an array of auditory filters; the processes that influence the masking of one sound by another; the range of sound levels that can be processed by the auditory system; the perception and modeling of loudness; level discrimination; the temporal resolution of the auditory system (the ability to detect changes over time); the perception and modeling of pitch for pure and complex tones; the perception of timbre for steady and time-varying sounds; the perception of space and sound localization; and the mechanisms underlying *auditory scene analysis* that allow the construction of percepts corresponding to individual sounds sources when listening to complex mixtures of sounds.

13.1 Absolute Thresholds

The absolute threshold of a sound is the lowest detectable level of that sound in the absence of any other sounds. In practice, there is no distinct sound level at which a sound suddenly becomes detectable. Rather, the probability of detecting a sound increases progressively as the sound level is increased from a very low value. Hence, the absolute threshold is defined as the sound level at which an individual detects the sound with a certain probability, such as 75% (in a two-alternative forced-choice task, where guessing leads to 50% correct, on average). Typically, results are averaged across many listeners with normal hearing (i. e., with no known history of hearing disorders and no obvious signs of hearing problems) to obtain representative results.

The absolute threshold for detecting sinusoids is partly determined by the sound transmission through the outer and middle ear (see Chap. 12); to a first approximation, the inner ear (the cochlea) is equally sensitive to all frequencies, except perhaps at very low frequencies and very high frequencies [13.3, 4]. Figure 13.1 shows estimates of the absolute threshold, measured in two ways. For the curve labeled MAP, standing for minimum audible pressure, the sound level was measured at a point close to the eardrum [13.2]. For the curve labeled MAF, standing for minimum audible field, the measurement of sound level was made after the listener had been removed from the sound field, at the point which had been occupied by the center of the listener's head [13.1]. For the MAF curve, the sound was presented in free field (in an anechoic chamber) from a direction directly in front of the listener. Note that the MAP estimates are for monaural listening and the MAF estimates are for binaural listening. On average, thresholds are about 2 dB lower when two ears are used as opposed to one, although the exact value of the difference varies across studies from 0 to 3 dB, and it can depend on the interaural phase of the tone [13.5–7]. Both curves represent the average data from many young listeners with normal hearing. It should be noted, however, that individual people may have thresholds as much as 20 dB above or below the mean at a specific frequency and still be considered as normal.

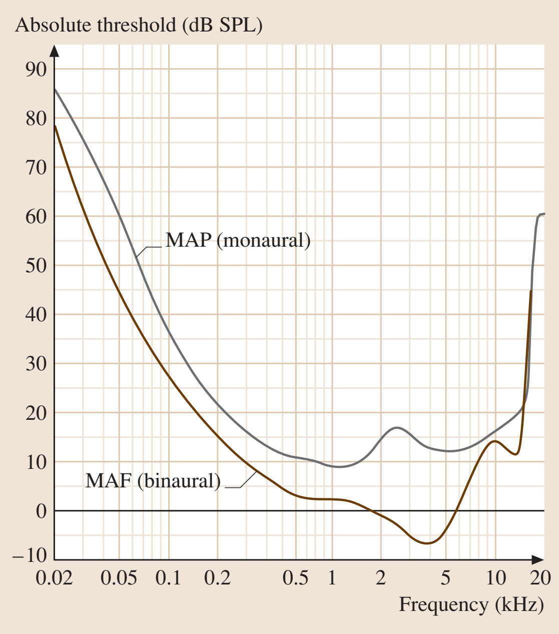

Fig. 13.1 The minimum audible sound level as a function of frequency. The *solid curve* shows the minimum audible field (MAF) for binaural listening published in an International Standards Organization (ISO) standard [13.1]. The *dashed curve* shows the minimum audible pressure (MAP) for monaural listening [13.2]

The MAP and MAF curves are somewhat differently shaped, since the head, the pinna and the meatus have an influence on the sound field. The MAP curve shows only minor peaks and dips (±5 dB) for frequencies between about 0.2 kHz and 13 kHz, whereas the MAF curve shows a distinct dip around 3–4 kHz and a peak around 8–9 kHz. The difference derives mainly from a broad resonance produced by the meatus and pinna. The sound level at the eardrum is enhanced markedly for frequencies in the range 1.5–6 kHz, with a maximum enhancement at 3 kHz of about 15 dB.

The highest audible frequency varies considerably with age. Young children can often hear tones as high as 20 kHz, but for most adults the threshold rises rapidly above about 15 kHz. The loss of sensitivity with increasing age (presbyacusis) is much greater at high frequencies than at low, and the variability between different people is also greater at high frequencies. There seems to be no well-defined low-frequency limit to hearing, although very intense low-frequency tones can sometimes be felt as vibration before they are heard. The low-frequency limit for the true hearing of pure tones probably lies at about 20 Hz. This is close to the lowest frequency which evokes a pitch sensation [13.8].

A third method of specifying absolute thresholds is commonly used when measuring hearing in clinical situations, for example, when a hearing impairment is

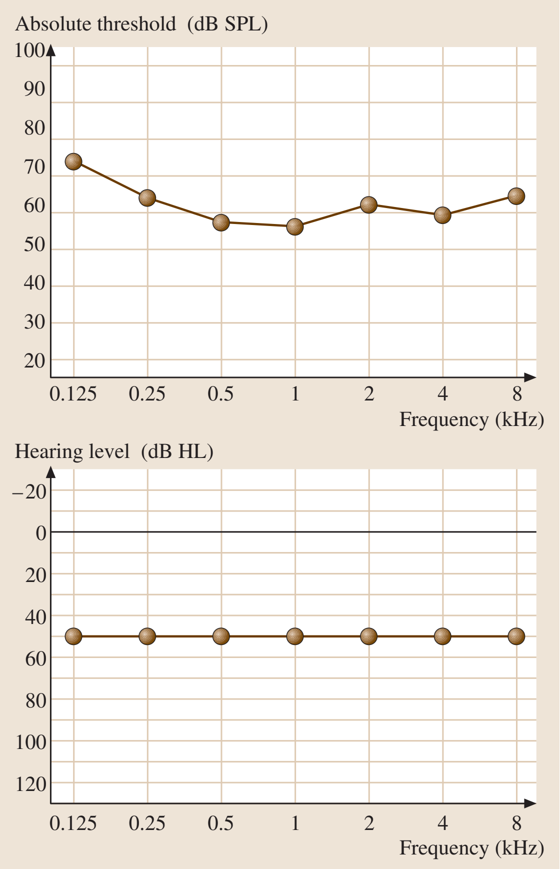

Fig. 13.2 Comparison of a clinical audiogram for a 50 dB hearing loss at all frequencies *(bottom)* and the absolute threshold curve for the same hearing loss plotted in terms of the MAP *(top)*

suspected; thresholds are specified relative to the average threshold at each frequency for young, healthy listeners with normal hearing. In this case, the sound level is usually specified relative to standardized values produced by a specific earphone in a specific coupler. A coupler is a device which contains a cavity or series of cavities and a microphone for measuring the sound produced by the earphone. The preferred earphone varies from one country to another. For example, the Telephonics TDH49 or TDH50 is often used in the UK and USA, while the Beyer DT48 is used in Germany. Thresholds specified in this way have units of dB HL (hearing level) in Europe or dB HTL (hearing threshold level) in the USA. For example, a threshold of 40 dB HL at 1 kHz would mean that the person had a threshold which was 40 dB higher than normal at that frequency. In psychoacoustic work, thresholds are normally plotted with threshold increasing upward, as in Fig. 13.1. However, in audiology, threshold elevations are shown as hearing losses, plotted downward. The average normal threshold is represented as a horizontal line at the top of the plot, and the degree of hearing loss is indicated by how much the threshold falls below this line. This type of plot is called an audiogram. Figure 13.2 compares an audiogram for a hypothetical hearing-impaired person with a flat hearing loss, with a plot of the same thresholds expressed as MAP values. Notice that, although the audiogram is flat, the corresponding MAP curve is not flat. Note also that thresholds in dB HL can be negative. For example, a threshold of −10 dB simply means that the individual is 10 dB more sensitive than the average.

The absolute thresholds described above were measured using tone bursts of relatively long duration. For durations exceeding about 500 ms, the sound level at threshold is roughly independent of duration. However, for durations less than about 200 ms, the sound level necessary for detection increases as duration decreases, by about 3 dB per halving of the duration [13.9]. Thus, the sound energy required for threshold is roughly constant.

13.2 Frequency Selectivity and Masking

Frequency selectivity refers to the ability to resolve or separate the sinusoidal components in a complex sound. It is a key aspect of the analysis of sounds by the auditory system, and it influences many aspects of auditory perception, including the perception of loudness, pitch and timbre. It is often demonstrated and measured by studying masking, which has been defined as 'The process by which the threshold of audibility for one sound is raised by the presence of another (masking) sound' [13.10]. It has been known for many years that a signal is most easily masked by a sound having frequency components close to, or the same as, those of the signal [13.11]. This led to the idea that our ability to separate the components of a complex sound depends, at least in part, on the frequency analysis that takes place on the basilar membrane (see Chap. 12).

13.2.1 The Concept of the Auditory Filter

Fletcher [13.13], following *Helmholtz* [13.14], suggested that the peripheral auditory system behaves as if it contains a bank of bandpass filters, with overlapping passbands. These filters are now called the auditory filters. Fletcher thought that the basilar membrane provided the basis for the auditory filters. Each location on the basilar membrane responds to a limited range of frequencies, so each different point corresponds to a filter with a different center frequency. When trying to detect a signal in a broadband noise background, the listener is assumed to make use of a filter with a center frequency close to that of the signal. This filter passes the signal but removes a great deal of the noise. Only the components in the noise which pass through the filter have any effect in masking the signal. It is usually assumed that the threshold for detecting the signal is determined by the amount of noise passing through the auditory filter; specifically, threshold is assumed to correspond to a certain signal-to-noise ratio at the output of the filter. This set of assumptions has come to be known as the *power spectrum model* of masking [13.15], since the stimuli are represented by their long-term power spectra, i. e., the short-term fluctuations in the masker are ignored.

The question considered next is: What is the shape of the auditory filter? In other words, how does its relative response change as a function of the input frequency? Most methods for estimating the shape of the auditory filter at a given center frequency are based on the assumptions of the power spectrum model of masking. The threshold of a signal whose frequency is fixed is measured in the presence of a masker whose spectral content is varied. It is assumed, as a first approximation, that the signal is detected using the single auditory filter which is centered on the frequency of the signal, and that threshold corresponds to a constant signal-to-masker ratio at the output of that filter. The methods described below both measure the shape of the filter using this technique.

13.2.2 Psychophysical Tuning Curves

One method of measuring the shape of the filter involves a procedure which is analogous in many ways to the determination of a neural tuning curve (see Chap. 12), and the resulting function is often called a psychophysical tuning curve (PTC). To determine a PTC, the signal is fixed in level, usually at a very low level, say, 10 dB above absolute threshold (called 10 dB sensation level, SL). The masker can be either a sinusoid or a band of noise covering a small frequency range.

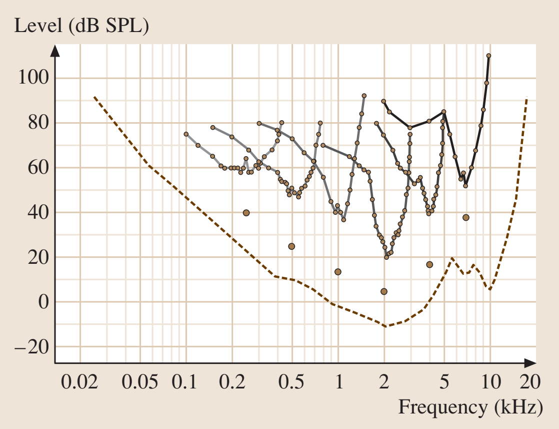

Fig. 13.3 Psychophysical tuning curves (PTCs) determined in simultaneous masking, using sinusoidal signals at 10 dB SL. For each curve, the *circle* below it indicates the frequency and level of the signal. The masker was a sinusoid which had a fixed starting phase relationship to the 50 ms signal. The masker level required for threshold is plotted as a function of masker frequency on a logarithmic scale. The *dashed line* shows the absolute threshold for the signal. (After *Vogten* [13.12])

For each of several masker frequencies, the level of the masker needed just to mask the signal is determined. Because the signal is at a low level it is assumed that it produces activity primarily at the output of a single auditory filter. It is assumed further that at threshold the masker produces a constant output from that filter, in order to mask the fixed signal. Thus the PTC indicates the masker level required to produce a fixed output from the auditory filter as a function of frequency. Normally a filter characteristic is determined by plotting the output from the filter for an input varying in frequency and fixed in level. However, if the filter is linear the two methods give the same result. Thus, assuming linearity, the shape of the auditory filter can be obtained simply by inverting the PTC. Examples of some PTCs are given in Fig. 13.3.

One problem in interpreting PTCs is that, in practice, the listener may use the information from more than one auditory filter. When the masker frequency is above the signal frequency the listener might do better to use the information from a filter centered just below the signal frequency. If the filter has a relatively flat top, and sloping edges, this will considerably attenuate the masker at the filter output, while only slightly attenuating the signal. By using this off-center filter the listener can improve performance. This is known as *off-frequency listening*, and there is now good evidence that humans do indeed listen off-frequency when it is advantageous to do so [13.16, 17]. The result of off-frequency listening is that the PTC has a sharper tip than would be obtained if only one auditory filter were involved [13.18].

13.2.3 The Notched-Noise Method

Patterson [13.19] described a method of determining auditory filter shape which limits off-frequency listening. The method is illustrated in Fig. 13.4. The signal (indicated by the bold vertical line) is fixed in frequency, and the masker is a noise with a bandstop or notch centered at the signal frequency. The deviation of each edge of the noise from the center frequency is denoted by Δf. The width of the notch is varied, and the threshold of the signal is determined as a function of notch width. Since the notch is symmetrically placed around the signal frequency, the method cannot reveal asymmetries in the auditory filter, and the analysis assumes that the filter is symmetric on a linear frequency scale. This assumption appears not unreasonable, at least for the top part of the filter and at moderate sound levels, since PTCs are quite symmetric around the tips. For a signal symmetrically placed in a bandstop noise, the optimum signal-to-masker ratio at the output of the auditory filter is achieved with a filter centered at the signal frequency, as illustrated in Fig. 13.4.

As the width of the spectral notch is increased, less noise passes through the auditory filter. Thus the threshold of the signal drops. The amount of noise passing through the auditory filter is proportional to the area under the filter in the frequency range covered by the noise. This is shown as the dark shaded areas in Fig. 13.4. Assuming that threshold corresponds to a constant signal-to-masker ratio at the output of the filter, the change in signal threshold with notch width indicates how the area under the filter varies with Δf. The area under a function between certain limits is obtained by integrating the value of the function over those limits. Hence by differentiating the function relating threshold to Δf, the relative response of the filter at that value of Δf is obtained. In other words, the relative response of the filter for a given deviation, Δf, from the center frequency is equal to the slope of the function relating signal threshold to notch width, at that value of Δf.

A typical auditory filter derived using this method is shown in Fig. 13.5. It has a rounded top and quite steep skirts. The sharpness of the filter is often specified as the bandwidth of the filter at which the response has fallen by a factor of two in power, i.e., by 3 dB. The 3 dB bandwidths of the auditory filters derived using the notched-noise method are typically between 10% and

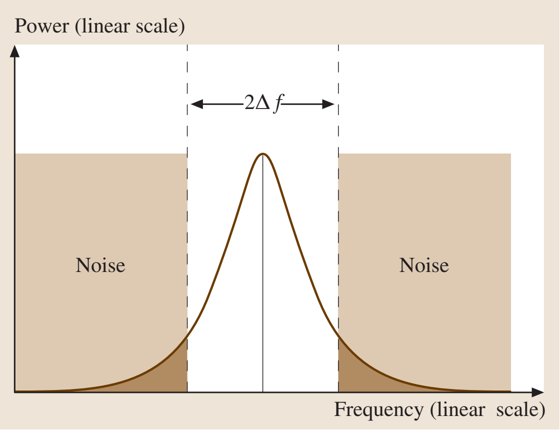

Fig. 13.4 Schematic illustration of the technique used by *Patterson*. [13.19] to determine the shape of the auditory filter. The threshold of the sinusoidal signal (indicated by the *bold vertical line*) is measured as a function of the width of a spectral notch in the noise masker. The amount of noise passing through the auditory filter centered at the signal frequency is proportional to the shaded areas

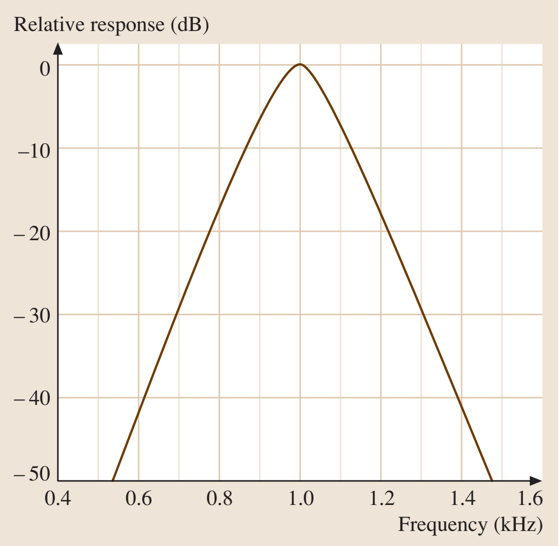

Fig. 13.5 A typical auditory filter shape determined using Patterson's method. The filter is centered at 1 kHz. The relative response of the filter (in dB) is plotted as a function of frequency

15% of the center frequency. An alternative measure is the equivalent rectangular bandwidth (ERB), which is the bandwidth of a rectangular filter which has the same peak transmission as the filter of interest and which passes the same total power for a white noise input. The ERB of the auditory filter is a little larger than the 3 dB bandwidth. In what follows, the mean ERB of the auditory filter determined using young listeners with normal hearing and using a moderate noise level is denoted $\mathrm{ERB_N}$ (where the subscript N denotes normal hearing). An equation describing the value of $\mathrm{ERB_N}$ as a function of center frequency, F (in Hz), is [13.20]:

$$\mathrm{ERB_N} = 24.7(0.00437F + 1)\,. \tag{13.1}$$

Sometimes it is useful to plot psychoacoustical data on a frequency scale related to $\mathrm{ERB_N}$. Essentially, the value of $\mathrm{ERB_N}$ is used as the unit of frequency. For example, the value of $\mathrm{ERB_N}$ for a center frequency of 1 kHz is about 132 Hz, so an increase in frequency from 934 to 1066 Hz represents a step of one $\mathrm{ERB_N}$. A formula relating $\mathrm{ERB_N}$ number to frequency is [13.20]:

$$\mathrm{ERB_N}\ \text{number} = 21.4\log_{10}(0.00437F + 1)\,, \tag{13.2}$$

where F is frequency in Hz. This scale is conceptually similar to the Bark scale proposed by *Zwicker* and coworkers [13.22], although it differs somewhat in numerical values.

The notched-noise method has been extended to include conditions where the spectral notch in the noise is placed asymmetrically about the signal frequency. This allows the measurement of any asymmetry in the auditory filter, but the analysis of the results is more difficult, and has to take off-frequency listening into account [13.23]. It is beyond the scope of this chapter to give details of the method of analysis; the interested reader is referred to [13.15, 20, 24, 25]. The results show that the auditory filter is reasonably symmetric at moderate sound levels, but becomes increasingly asymmetric at high levels, the low-frequency side becoming shallower than the high-frequency side. The filter shapes derived using the notched-noise method are quite similar to inverted PTCs [13.26], except that PTCs are slightly sharper around their tips, probably as a result of off-frequency listening.

13.2.4 Masking Patterns and Excitation Patterns

In the masking experiments described so far, the frequency of the signal was held constant, while the masker was varied. These experiments are most appropriate for estimating the shape of the auditory filter at a given center frequency. However, many of the early experiments on masking did the opposite; the masker was held constant in both level and frequency and the signal threshold was measured as a function of the signal frequency. The resulting functions are called masking patterns or masked audiograms.

Masking patterns show steep slopes on the low-frequency side (when the signal frequency is below that of the masker), of between 55 and 240 dB/octave. The slopes on the high-frequency side are less steep and depend on the level of the masker. A typical set of results is shown in Fig. 13.6. Notice that on the high-frequency side the curve is shallower at the highest level. Around the tip of the masking pattern, the growth of masking is approximately linear; a 10 dB increase in masker level leads to roughly a 10 dB increase in the signal threshold. However, for signal frequencies in the range from about 1300 to 2000 Hz, when the level of the masker is increased by 10 dB (e.g., from 70 to 80 dB SPL), the masked threshold increases by more than 10 dB; the amount of masking grows nonlinearly on the high-frequency side. This has been called the *upward spread of masking*.

The masking patterns do not reflect the use of a single auditory filter. Rather, for each signal frequency the listener uses a filter centered close to the signal frequency. Thus the auditory filter is shifted as the signal frequency is altered. One way of interpreting the masking pattern

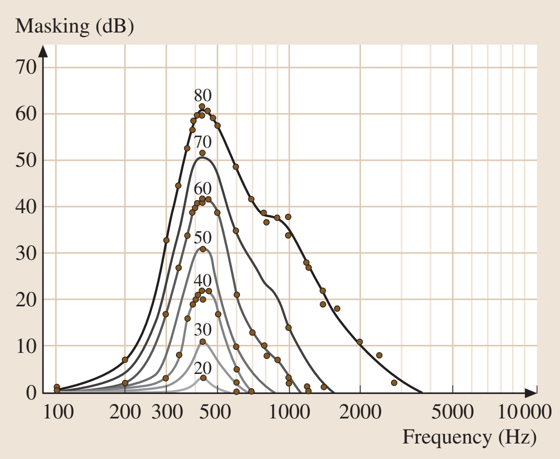

Fig. 13.6 Masking patterns for a narrow-band noise masker centered at 410 Hz. Each curve shows the elevation in threshold of a pure-tone signal as a function of signal frequency. The overall noise level for each curve is indicated in the figure (After *Egan* and *Hake* [13.21])

is as a crude indicator of the excitation pattern of the masker [13.27]. The excitation pattern is a representation of the effective amount of excitation produced by a stimulus as a function of characteristic frequency (CF) on the basilar membrane (BM; see Chap. 12), and is plotted as effective level (in dB) against CF. In the case of a masking sound, the excitation pattern can be thought of as representing the relative amount of vibration produced by the masker at different places along the basilar membrane. The signal is detected when the excitation it produces is some constant proportion of the excitation produced by the masker at places with CFs close to the signal frequency. Thus the threshold of the signal as a function of frequency is proportional to the masker excitation level. The masking pattern should be parallel to the excitation pattern of the masker, but shifted vertically by a small amount. In practice, the situation is not so straightforward, since the shape of the masking pattern is influenced by factors such as off-frequency listening, and the detection of beats and combination tones [13.28].

Moore and *Glasberg* [13.29] have described a way of deriving the shapes of excitation patterns using the concept of the auditory filter. They suggested that the excitation pattern of a given sound can be thought of as the output of the auditory filters plotted as a function of their center frequency. To calculate the excitation pattern of a given sound, it is necessary to calculate the output of each auditory filter in response to that sound, and to plot the output as a function of the filter center frequency. The characteristics of the auditory filters are determined using the notched-noise method described earlier. Figure 13.7 shows excitation patterns calculated in this way for 1000 Hz sinusoids with various levels. The patterns are similar in form to the masking patterns shown in Fig. 13.6. Software for calculating excitation patterns can be downloaded from http://hearing.psychol.cam.ac.uk/Demos/demos.html.

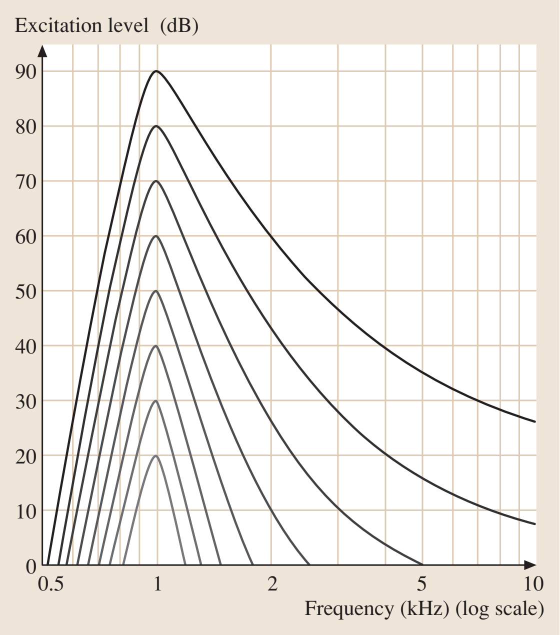

Fig. 13.7 Calculated psychoacoustical excitation patterns for a 1 kHz sinusoid at levels ranging from 20 to 90 dB SPL in 10 dB steps

13.2.5 Forward Masking

Masking can occur when the signal is presented just before or after the masker. This is called non-simultaneous masking and it is studied using brief signals, often called *probes*. Two basic types of non-simultaneous masking can be distinguished: (1) backward masking, in which the probe precedes the masker; and (2) forward masking, in which the probe follows the masker. Although many studies of backward masking have been published, the phenomenon is poorly understood. The amount of backward masking obtained depends strongly on how much practice the subjects have received, and practiced subjects often show little or no backward masking [13.30, 31]. The larger masking effects found for unpracticed subjects may reflect some sort of confusion of the signal with the masker. In the following, I will focus on forward masking, which can be substantial even in highly practiced subjects. The main properties of forward masking are as follows:

1. Forward masking is greater the nearer in time to the masker that the signal occurs. This is illustrated in the left panel of Fig. 13.8. When the delay D of the signal after the end of the masker is plotted on a logarithmic scale, the data fall roughly on a straight line. In other words, the amount of forward masking, in dB, is a linear function of $\log D$.
2. The rate of recovery from forward masking is greater for higher masker levels. Regardless of the initial amount of forward masking, the masking decays to zero after 100–200 ms.
3. A given increment in masker level does not produce an equal increment in amount of forward masking. For example, if the masker level is increased by 10 dB, the masked threshold may only increase by 3 dB. This contrasts with simultaneous mask-

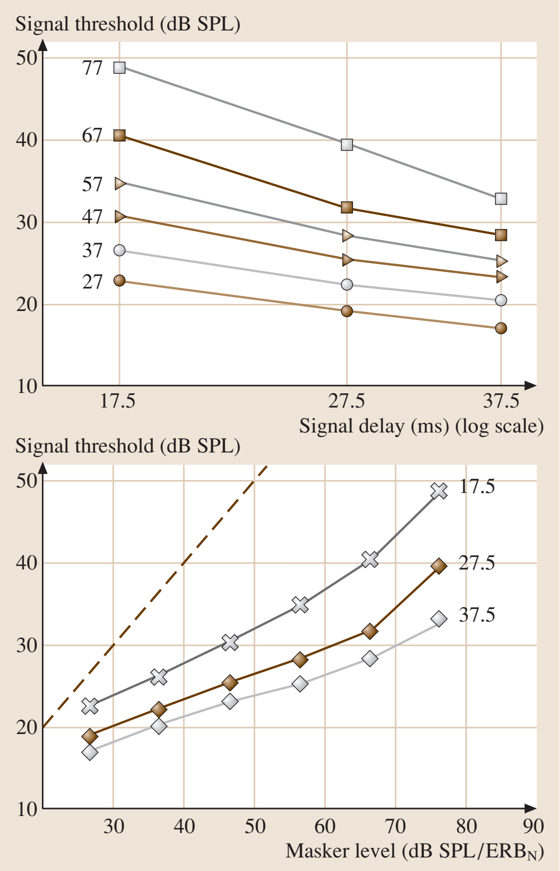

Fig. 13.8 The top panel shows the amount of forward masking of a brief 4 kHz signal, plotted as a function of the time delay of the signal after the end of the noise masker. Each curve shows results for a different noise level, specified as the level in a one-$\mathrm{ERB_N}$-wide band centered at 4 kHz. The results for each level fall roughly on a straight line when the signal delay is plotted on a logarithmic scale, as here. The bottom panel shows the same thresholds plotted as a function of masker level. Each curve shows results for a different signal delay time (17.5, 27.5, or 37.5 ms). Note that the slopes of these growth of masking functions decrease with increasing signal delay. The *dashed line* indicates a slope of 1. (After *Moore* and *Glasberg* [13.32])

ing, where, at least for wide-band maskers, threshold corresponds approximately to a constant signal-to-masker ratio. This effect can be quantified by plotting the signal threshold as a function of the masker level. The resulting function is called a growth of masking function. Several such functions are shown in the right panel of Fig. 13.8. In simultaneous masking such functions would have slopes close to one, as indicated by the dashed line. In forward masking the slopes are usually less than one, and the slopes decrease as the value of D increases.

4. The amount of forward masking increases with increasing masker duration for durations up to at least 50 ms. The results for greater masker durations vary somewhat across studies. Some studies show an effect of masker duration for durations up to 200 ms [13.33, 34], while others show little effect for durations beyond 50 ms [13.35].
5. Forward masking is influenced by the relation between the frequencies of the signal and the masker (just as in the case of simultaneous masking).

The basis of forward masking is still not clear. Four factors may contribute:

1. The response of the BM to the masker continues for some time after the end of the masker, an effect known as *ringing*. If the ringing overlaps with the response to the signal, then this may contribute to the masking of the signal. The duration of the ringing is less at places tuned to high frequencies, whose bandwidth is larger than at low frequencies. Hence, ringing plays a significant role only at low frequencies [13.36–38]. For frequencies above 200–300 Hz, the amount of forward masking for a given masker level and signal delay time is roughly independent of frequency [13.39].
2. The masker produces short-term adaptation or fatigue in the auditory nerve or at higher centers in the auditory system, which reduces the response to a signal presented just after the end of the masker [13.40]. However, the effect in the auditory nerve appears to be too small to account for the forward masking observed behaviorally [13.41].
3. The neural activity evoked by the masker persists at some level in the auditory system higher than the auditory nerve, and this persisting activity masks the signal [13.42].
4. The masker may evoke a form of inhibition in the central auditory system, and this inhibition persists for some time after the end of the masker [13.43].

Oxenham and *Moore* [13.44] have suggested that the shallow slopes of the growth of masking functions, as shown in the bottom panel of Fig. 13.8, can be explained, at least qualitatively, in terms of the compressive input–output function of the BM (see Chap. 12). Such an input–output function is shown schematically

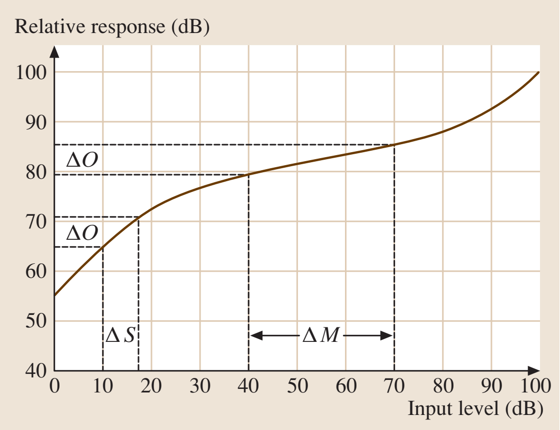

Fig. 13.9 Illustration of why growth of masking functions in forward masking usually have shallow slopes. The *solid curve* shows a schematic input–output function on the basilar membrane. The relative response is plotted on a dB scale with an arbitrary origin. When the masker is increased in level by ΔM, this produces an increase in response of ΔO. To restore signal threshold, the response to the signal also has to be increased by ΔO. This requires an increase in signal level, ΔS, which is markedly smaller than ΔM

in Fig. 13.9. It has a shallow slope for medium input levels, but a steeper slope at very low input levels. Assume that, for a given time delay of the signal relative to the masker, the response evoked by the signal at threshold is directly proportional to the response evoked by the masker. Assume, as an example, that a masker with a level of 40 dB produces a signal threshold of 10 dB. Consider now what happens when the masker level is increased by 30 dB. The increase in masker level, denoted by ΔM in Fig. 13.9, produces a relatively small increase in response, ΔO. To restore the signal to threshold, the signal has to be increased in level so that the response to it also increases by ΔO. However, this requires a relatively small increase in signal level, ΔS, as the signal level falls in the range where the input–output function is relatively steep. Thus, the growth of masking function has a shallow slope.

According to this explanation, the shallow slope of the growth of masking function arises from the fact that the signal level is lower than the masker level, so the masker is subject to more compression than the signal. The input–output function on the BM has a slope which decreases progressively with increasing level over the range 0 to about 50 dB. Hence the slope of the growth of masking function should decrease with increasing difference in level between the masker and signal. This can account for the progressive decrease in the slopes of the growth of masking functions with increasing delay between the signal and masker (see the right-hand panel of Fig. 13.8); longer delays are associated with greater differences in level between the signal and masker. Another prediction is that the growth of masking function for a given signal delay should increase in slope if the signal level is high enough to fall in the compressive region of the input–output function. Such an effect can be seen in the growth of masking function for the shortest delay time in Fig. 13.8; the function steepens for the highest signal level.

In summary, the processes underlying forward masking are not fully understood. Contributions from a number of different sources may be important. Temporal overlap of patterns of vibration on the BM may be important, especially for small delay times between the signal and masker. Short-term adaptation or fatigue in the auditory nerve may also play a role. At higher neural levels, a persistence of the excitation or inhibition evoked by the masker may occur. The form of the growth of masking functions can be explained, at least qualitatively, in terms of the nonlinear input–output functions observed on the BM.

13.2.6 Hearing Out Partials in Complex Tones

A complex tone, such as a tone produced by a musical instrument, usually evokes a single pitch; pitches corresponding to the frequencies of individual partials are not usually perceived. However, such pitches can be heard if attention is directed appropriately. In other words, individual partials can be *heard out*. *Plomp* [13.45] and *Plomp* and *Mimpen* [13.46] used complex tones with 12 equal-amplitude sinusoidal components to investigate the limits of this ability. The listener was presented with two comparison tones, one of which was of the same frequency as a partial in the complex; the other lay halfway between that frequency and the frequency of the adjacent higher or lower partial. The listener had to judge which of these two tones was a component of the complex. Plomp used two types of complex: a harmonic complex containing harmonics 1 to 12, where the frequencies of the components were integer multiples of that of the fundamental; and a nonharmonic complex, where the frequencies of the components were mistuned from simple frequency ratios. He found that for both kinds of complex, partials could only be heard out if they were sufficiently far in frequency from neighboring partials. The data, and other more recent data [13.47], are consistent with the hypothesis that a partial can be heard out (with 75% accuracy) when it is separated from neighboring (equal-amplitude) partials by 1.25 times the $\mathrm{ERB_N}$ of the auditory filter. For harmonic complex tones, only the first (lowest) five to eight harmonics can be heard out, as higher harmonics are separated by less than 1.25 $\mathrm{ERB_N}$.

It seems likely that the analysis of partials from a complex sound depends in part on factors other than the frequency analysis that takes place on the basilar membrane. *Soderquist* [13.48] compared musicians and non-musicians in a task very similar to that of Plomp, and found that the musicians were markedly superior. This result could mean that musicians have smaller auditory filter bandwidths than non-musicians. However, *Fine* and *Moore* [13.49] showed that auditory filter bandwidths, as estimated in a notched-noise masking experiment, did not differ for musicians and non-musicians. It seems that some mechanism other than peripheral filtering is involved in hearing out partials from complex tones and that musicians, because of their greater experience, are able to make more efficient use of this mechanism.

13.3 Loudness

The human ear is remarkable both in terms of its absolute sensitivity and the range of sound intensities to which it can respond. The most intense sound that can be heard without damaging the ear has a level about 120 dB above the absolute threshold; this range is referred to as the dynamic range of the auditory system and it corresponds to a ratio of intensities of 1 000 000 000 000:1.

Loudness corresponds to the subjective impression of the magnitude of a sound. The formal definition of loudness is: that intensive attribute of auditory sensation in terms of which sounds can be ordered on a scale extending from soft to loud [13.10]. Because loudness is subjective, it is very difficult to measure in a quantitative way. Estimates of loudness can be strongly affected by bias and context effects of various kinds [13.50, 51]. For example the impression of loudness of a sound with a moderate level (say, 60 dB SPL) can be affected by presenting a high level sound (say, 100 dB SPL) before the moderate level sound.

13.3.1 Loudness Level and Equal-Loudness Contours

It is often useful to be able to compare the loudness of sounds with that of a standard, reference sound. The most common reference sound is a 1000 Hz sinusoid, presented binaurally in a free field, with the sound coming from directly in front of the listener. The loudness level of a sound is defined as the intensity level of a 1000 Hz sinusoid that is equal in loudness to the sound. The unit of loudness level is the phon. Thus, the loudness level of any sound in phons is the level (in dB SPL) of the 1000 Hz sinusoid to which it sounds equal in loudness. For example, if a sound appears to be as loud as a 1000 Hz sinusoid with a level of 45 dB SPL, then the sound has a loudness level of 45 phons. To determine the loudness level of a given sound, the subject is asked to adjust the level of a 1000 Hz sinusoid until it appears to have the same loudness as that sound. The 1000 Hz sinusoid and the test sound are presented alternately rather than simultaneously.

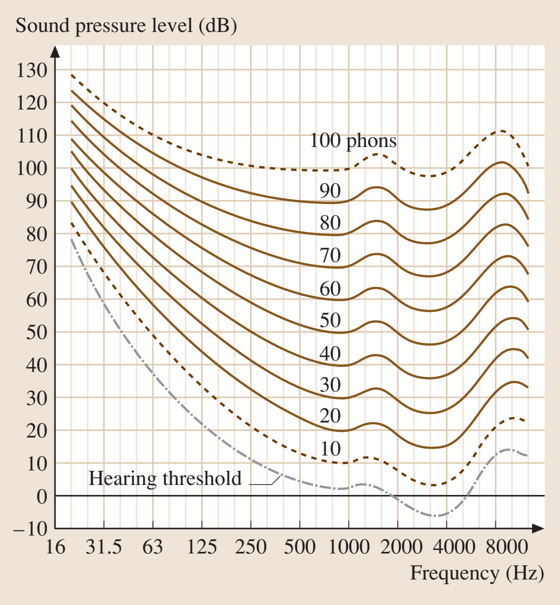

Fig. 13.10 Equal-loudness contours for loudness levels from 10 to 100 phons for sounds presented binaurally from the frontal direction. The absolute threshold curve (the MAF) is also shown. The curves for loudness levels of 10 and 100 phons are *dashed*, as they are based on interpolation and extrapolation, respectively

In a variation of this procedure, the 1000 Hz sinusoid is fixed in level, and the test sound is adjusted to give a loudness match, again with alternating presentation. If this is repeated for various different frequencies of a sinusoidal test sound, an equal-loudness contour is generated [13.52, 53]. For example, if the 1000 Hz sinusoid is fixed in level at 40 dB SPL, then the 40 phon equal-loudness contour is generated. Figure 13.10 shows equal-loudness contours as published in a recent standard [13.53]. The figure shows equal-loudness contours for binaural listening for loudness levels from 10–100 phons, and it also includes the absolute threshold (MAF) curve. The listening conditions were similar to those for determining the MAF curve; the sound came from a frontal direction in a free field. The equal-loudness contours are of similar shape to the MAF curve, but tend to become flatter at high loudness levels. Note that the MAF curve in Fig. 13.10 differs somewhat from that in Fig. 13.1, as the two curves are based on different standards.

Note that the subjective loudness of a sound is not directly proportional to its loudness level in phons. For example, a sound with a loudness level of 80 phons sounds much more than twice as loud as a sound with a loudness level of 40 phons. This is discussed in more detail in the next section.

13.3.2 The Scaling of Loudness

Several methods have been developed that attempt to measure directly the relationship between the physical magnitude of sound and perceived loudness [13.54]. In one, called magnitude estimation, sounds with various different levels are presented, and the subject is asked to assign a number to each one according to its perceived loudness. In a second method, called magnitude production, the subject is asked to adjust the level of a sound until it has a loudness corresponding to a specified number.

On the basis of results from these two methods, Stevens suggested that loudness, L, was a power function of physical intensity, I:

$$L = kI^{0.3}x\ , \tag{13.3}$$

where k is a constant depending on the subject and the units used. In other words, the loudness of a given sound is proportional to its intensity raised to the power 0.3. Note that this implies that loudness is *not* linearly related to intensity; rather, it is a *compressive* function of intensity. An approximation to this equation is that the loudness doubles when the intensity is increased by a factor of 10, or, equivalently, when the level is increased by 10 dB. In practice, this relationship only holds for sound levels above about 40 dB SPL. For lower levels than this, the loudness changes with intensity more rapidly than predicted by the power-law equation.

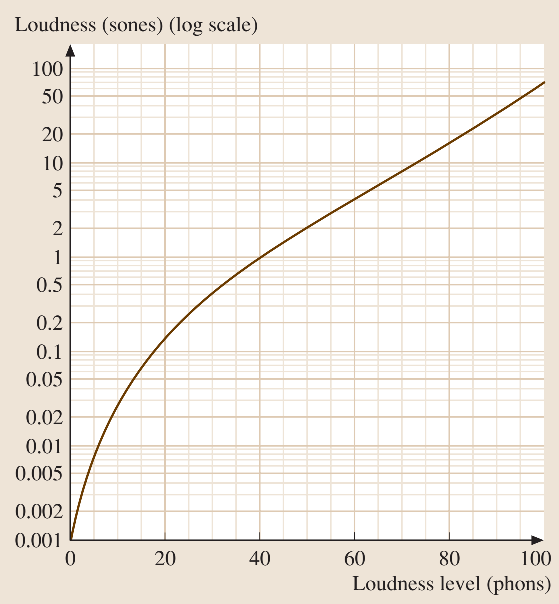

Fig. 13.11 The relationship between loudness in sones and loudness level in phons for a 1000 Hz sinusoid. The curve is based on the loudness model of *Moore* et al.[13.3]

The unit of loudness is the *sone*. One sone is defined arbitrarily as the loudness of a 1000 Hz sinusoid at 40 dB SPL, presented binaurally from a frontal direction in a free field. Fig. 13.11 shows the relationship between loudness in sones and the physical level of a 1000 Hz sinusoid, presented binaurally from a frontal direction in a free-field; the level of the 1000 Hz tone is equal to its loudness level in phons. This figure is based on predictions of a loudness model [13.3], but it is consistent with empirical data obtained using scaling methods [13.55]. Since the loudness in sones is plotted on a logarithmic scale, and the decibel scale is itself logarithmic, the curve shown in Fig. 13.11 approximates a straight line for levels above 40 dB SPL. The slope corresponds to a doubling of loudness for each 10 dB increase in sound level.

13.3.3 Neural Coding and Modeling of Loudness

The mechanisms underlying the perception of loudness are not fully understood. A common assumption is that

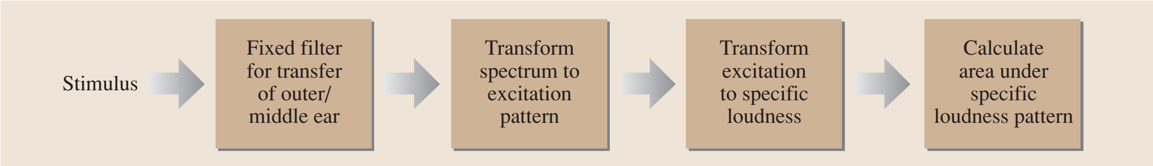
Stimulus
Fixed filter for transfer of outer/ middle ear
Transform spectrum to excitation pattern
Transform excitation to specific loudness
Calculate area under specific loudness pattern

Fig. 13.12 Block diagram of a typical loudness model

loudness is somehow related to the total neural activity evoked by a sound, although this concept has been questioned [13.56]. In any case, it is commonly assumed that loudness depends upon a summation of loudness contributions from different frequency channels (i. e., different auditory filters). Models incorporating this basic concept have been proposed by *Fletcher* and *Munson* [13.57], by *Zwicker* [13.58, 59] and by *Moore* and coworkers [13.3, 60]. The models attempt to calculate the average loudness that would be perceived by a large group of listeners with normal hearing under conditions where biases are minimized as far as possible. The models all have the basic form illustrated in Fig. 13.12. The first stage is a fixed filter to account for the transmission of sound through the outer and middle ear. The next stage is the calculation of an excitation pattern for the sound under consideration. In most of the models, the excitation pattern is calculated from psychoacoustical masking data, as described earlier. From the excitation-pattern stage onwards, the models should be considered as multichannel; the excitation pattern is sampled at regular intervals along the $\mathrm{ERB_N}$ number scale, each sample value corresponding to the amount of excitation at a specific center frequency.

The next stage is the transformation from excitation level (dB) to specific loudness, which is a kind of *loudness density*. It represents the loudness that would be evoked by the excitation within a small fixed distance along the basilar membrane if it were possible to present that excitation alone (without any excitation at adjacent regions on the basilar membrane). In the model of *Moore* et al. [13.3], the distance is 0.89 mm, which corresponds to one $\mathrm{ERB_N}$, so the specific loudness represents the loudness per $\mathrm{ERB_N}$. The specific loudness plotted as a function of $\mathrm{ERB_N}$ number is called the *specific loudness pattern*. The specific loudness cannot be measured either physically or subjectively. It is a theoretical construct used as an intermediate stage in the loudness models. The transformation from excitation level to specific loudness involves a compressive nonlinearity. Although the models are based on psychoacoustical data, this transformation can be thought of as representing the way that physical excitation is transformed into neural activity; the specific loudness is assumed to be related to the amount of neural activity at the corresponding CF. The overall loudness of a given sound, in sones, is assumed to be proportional to the total area under the specific loudness pattern. In other words, the overall loudness is calculated by summing the specific loudness values across all $\mathrm{ERB_N}$ numbers (corresponding to all center frequencies).

Loudness models of this type have been rather successful in accounting for experimental data on the loudness of both simple sounds and complex sounds [13.3]. They have also been incorporated into loudness meters, which can give an appropriate indication of the loudness of sounds even when they fluctuate over time [13.27, 60, 61]. Software for calculating loudness using the model of *Moore* et al. [13.3] can be downloaded from http://hearing.psychol.cam.ac.uk/Demos/demos.html. The new American National Standards Institute (ANSI) standard for calculating loudness [13.62] is based on this model.

13.3.4 The Effect of Bandwidth on Loudness

Consider a complex sound of fixed energy (or intensity) having a bandwidth W. If W is less than a certain bandwidth, called the critical bandwidth for loudness, $\mathrm{CB_L}$, then the loudness of the sound is almost independent of W; the sound is judged to be about as loud as a pure tone or narrow band of noise of equal intensity lying at the center frequency of the band. However, if W is increased beyond $\mathrm{CB_L}$, the loudness of the complex begins to increase. This has been found to be the case for bands of noise [13.27, 63] and for complex sounds consisting of several pure tones whose frequency separation is varied [13.64, 65]. An example for bands of noise is given in Fig. 13.13. The $\mathrm{CB_L}$ for the data in Fig. 13.13 is about 250–300 Hz for a center frequency of 1420 Hz, although the exact value of $\mathrm{CB_L}$ is hard to determine. The value of $\mathrm{CB_L}$ is similar to, but a little greater than, the $\mathrm{ERB_N}$ of the auditory filter. Thus, for a given amount of energy, a complex sound is louder if its bandwidth exceeds one $\mathrm{ERB_N}$ than if its bandwidth is less than one $\mathrm{ERB_N}$.

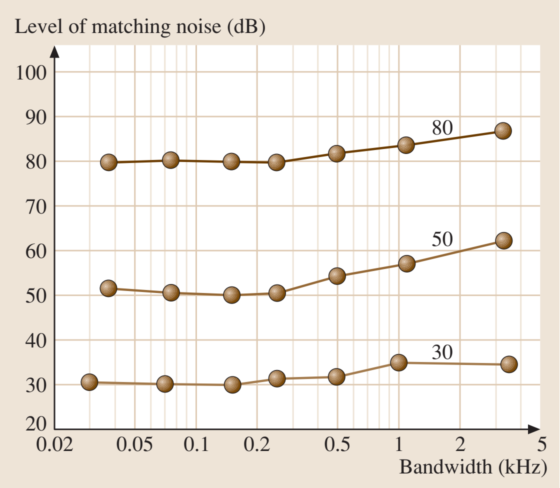

Fig. 13.13 The level of a 210 Hz-wide noise required to match the loudness of a noise of variable bandwidth is plotted as a function of the variable bandwidth; the bands of noise were geometrically centered at 1420 Hz. The number by each curve is the overall noise level in dB SPL. (After *Zwicker* et al.[13.63])

The pattern of results shown in Fig. 13.13 can be understood in terms of the loudness models described above. First, a qualitative account is given to illustrate the basic concept. When a sound has a bandwidth less than one ERB_N at a given center frequency, the excitation pattern and the specific loudness pattern are roughly independent of the bandwidth of the sound [13.66]. Further, the overall loudness is dominated by the specific loudness at the peak of the pattern. Consider now the effect of increasing the bandwidth of a sound from one ERB_N to N ERB_N, keeping the overall power, P_{overall}, constant. The power in each one-ERB_N-wide band is now P_{overall}/N. However, the peak specific loudness evoked by each band decreases by much less than a factor of N, because of the compressive relationship between excitation and specific loudness. For example, if $N = 10$, the peak specific loudness evoked by each band decreases by a factor of about two. Thus, in this example, there are ten bands each about half as loud as the original one-ERB_N-wide band, so the overall loudness is about a factor of five greater for the ten-ERB_N-wide than for the one-ERB_N-wide band. Generally, once the bandwidth is increased beyond one ERB_N, the loudness goes up because the decrease in loudness per ERB_N is more than offset by the increase in the number of channels contributing significantly to the overall loudness.

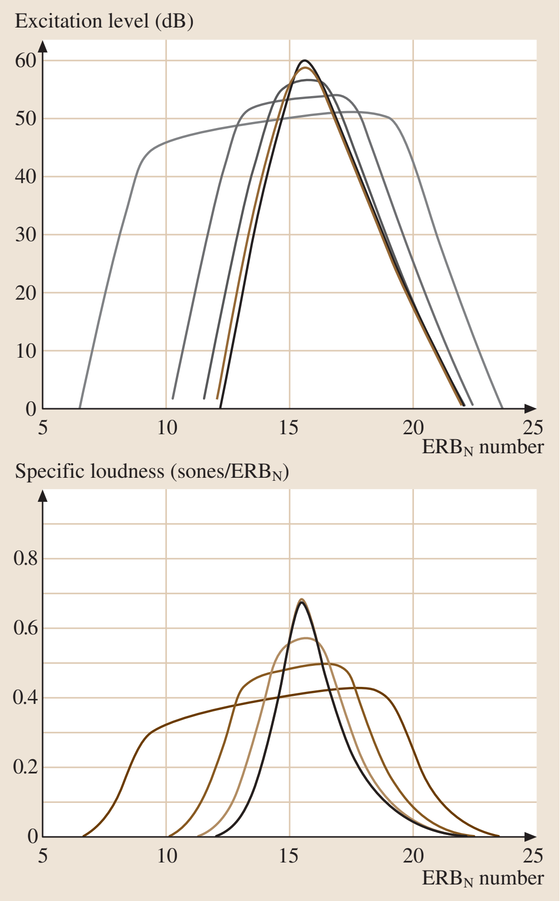

Fig. 13.14 The upper panel shows excitation patterns for a 1 kHz sinusoid with a level of 60 dB SPL (the narrowest pattern with the highest peak) and for noise bands of various widths, all centered at 1 kHz and with an overall level of 60 dB SPL. The frequency scale has been transformed to an ERB_N number scale. The noise bands have widths of 20, 60, 140, 300, 620 and 1260 Hz. As the bandwidth is increased, the patterns decrease in height but spread over a greater number of ERB_Ns. The lower panel shows specific loudness patterns corresponding to the excitation patterns in the upper panel. For bandwidths up to 140 Hz, the area under the specific loudness patterns is constant. For greater bandwidths, the total area increases

This argument is illustrated in a more quantitative way in Fig. 13.14, which shows excitation patterns (top) and specific loudness patterns (bottom) for a sinusoid and for bands of noise of various widths (20, 40, 60, 140, 300 and 620 Hz), all with a level of 60 dB SPL, calculated using the model of *Moore* et al. [13.3]. With increasing bandwidth, the specific loudness patterns become lower at their tips, but broader. For the first three bandwidths, the small decrease in area around the tip is almost exactly canceled by the increase on the skirts, so the total area remains almost constant. When the bandwidth is increased beyond 140 Hz, the increase on the skirts is greater than the decrease around the tip, and so the total area, and hence the predicted loudness, increases. The value of $\mathrm{CB_L}$ in this case is a little greater than 140 Hz. Since the increase in loudness depends on the summation of specific loudness at different center frequencies, the increase in loudness is often described as loudness summation.

At low sensation levels (around 10–20 dB SL), the loudness of a complex sound is roughly independent of bandwidth. This can also be explained in terms of the loudness models described above. At these low levels, specific loudness changes rapidly with excitation level, and so does loudness. As a result, the total area under the specific loudness pattern remains almost constant as the bandwidth is altered, even for bandwidths greater than $\mathrm{CB_L}$. Thus, loudness is independent of bandwidth. When a narrow-band sound has a very low sensation level (below 10 dB), then if the bandwidth is increased keeping the total energy constant, the output of each auditory filter becomes insufficient to make the sound audible. Accordingly, near threshold, loudness must decrease as the bandwidth of a complex sound is increased from a small value. As a consequence, if the intensity of a complex sound is increased slowly from a subthreshold value, the rate of growth of loudness is greater for a wide-band sound than for a narrow-band sound.

13.3.5 Intensity Discrimination

The smallest detectable change in intensity of a sound has been measured for many different types of stimuli by a variety of methods. The three main methods are:

1. Modulation detection. The stimulus is amplitude modulated (i. e., made to vary in amplitude) at a slow regular rate and the listener is required to detect the modulation. Usually, the modulation is sinusoidal.
2. Increment detection. A continuous background stimulus is presented, and the subject is required to detect a brief increment in the level of the background. Often the increment is presented at one of two possible times, indicated by lights, and the listener is required to indicate whether the increment occurred synchronously with the first light or the second light.
3. Intensity discrimination of gated or pulsed stimuli. Two (or more) separate pulses of sound are presented successively, one being more intense than the other(s), and the subject is required to indicate which pulse was the most intense.

In all of these tasks, the subjective impression of the listener is of a change in loudness. For example, in method 1 the modulation is heard as a fluctuation in loudness. In method 2 the increment is heard as a brief increase in loudness of the background, or sometimes as an extra sound superimposed on the background. In method 3, the most intense pulse appears louder than the other(s). Although there are some minor discrepancies in the experimental results for the different methods, the general trend is similar. For wide-band noise, or for bandpass-filtered noise, the smallest detectable intensity change, ΔI, is approximately a constant fraction of the intensity of the stimulus, I. In other words, $\Delta I/I$ is roughly constant. This is an example of Weber's Law, which states that the smallest detectable change in a stimulus is proportional to the magnitude of that stimulus. The value of $\Delta I/I$ is called the Weber fraction. Thresholds for detecting intensity changes are often specified as the change in level at threshold, ΔL, in decibels. The value of ΔL is given by

$$\Delta L = 10\log_{10}[(I+\Delta I)/I]\,. \tag{13.4}$$

As $\Delta I/I$ is constant, ΔL is also constant, regardless of the absolute level, and for wide-band noise has a value of about 0.5–1 dB. This holds from about 20 dB above threshold to 100 dB above threshold [13.67]. The value of ΔL increases for sounds which are close to the absolute threshold.

For sinusoidal stimuli, the situation is somewhat different. If ΔI (in dB) is plotted against I (also in dB), a straight line is obtained with a slope of about 0.9; Weber's law would give a slope of 1.0. Thus discrimination, as measured by the Weber fraction, improves at high levels. This has been termed the "near miss" to Weber's Law. The data of Riesz [13.68] for modulation detection show a value of ΔL of 1.5 dB at 20 dB SL, 0.7 dB at 40 dB SL, and 0.3 dB at 80 dB SL (all at 1000 Hz). The Weber fraction may increase somewhat at very high sound levels (above 100 dB SPL) [13.69]. In everyday life, a change in level of 1 dB would hardly be noticed, but a change in level of 3 dB (corresponding to a doubling or halving of intensity) would be fairly easily heard.

13.4 Temporal Processing in the Auditory System

This section is concerned mainly with temporal resolution (or acuity), which refers to the ability to detect changes in stimuli over time, for example, to detect a brief gap between two stimuli or to detect that a sound is modulated in some way. As pointed out by *Viemeister* and *Plack* [13.71], it is also important to distinguish the rapid pressure variations in a sound (the *fine structure*) from the slower overall changes in the amplitude of those fluctuations (the *envelope*). Temporal resolution normally refers to the resolution of changes in the envelope, not in the fine structure. In characterizing temporal resolution in the auditory system, it is important to take account of the filtering that takes place in the peripheral auditory system. Temporal resolution depends on two main processes: analysis of the time pattern occurring within each frequency channel and comparison of the time patterns across channels.

A major difficulty in measuring the temporal resolution of the auditory system is that changes in the time pattern of a sound are generally associated with changes in its magnitude spectrum, for example its power spectrum (see Chap. 14). Thus, the detection of a change in time pattern can sometimes depend not on temporal resolution per se, but on the detection of the change in magnitude spectrum. There have been two general approaches to getting around this problem. One is to use signals whose magnitude spectrum is not changed when the time pattern is altered. For example, the magnitude spectrum of white noise remains flat if the noise is interrupted, i. e., if a gap is introduced into the noise. The second approach uses stimuli whose magnitude spectra are altered by the change in time pattern, but extra background sounds are added to mask the spectral changes. Both approaches will be described.

13.4.1 Temporal Resolution Based on Within-Channel Processes

The threshold for detecting a gap in a broadband noise provides a simple and convenient measure of temporal resolution. Usually a two-alternative forced-choice (2AFC) procedure is used: the subject is presented with two successive bursts of noise and either the first or the second burst (at random) is interrupted to produce the gap. The task of the subject is to indicate which burst contained the gap. The gap threshold is typically 2–3 ms [13.72, 73]. The threshold increases at very low sound levels, when the level of the noise approaches the absolute threshold, but is relatively invariant with level for moderate to high levels.

The long-term magnitude spectrum of a sound is not changed when that sound is time reversed (played backward in time). Thus, if a time-reversed sound can be discriminated from the original, this must reflect a sensitivity to the difference in time pattern of the two sounds. This was exploited by *Ronken* [13.74], who used as stimuli pairs of clicks differing in amplitude. One click, labeled A, had an amplitude greater than that of the other click, labeled B. Typically the amplitude of A was twice that of B. Subjects were required to distinguish click pairs differing in the order of A and B: either AB or BA. The ability to do this was measured as a function of the time interval or gap between A and B. *Ronken* found that subjects could distinguish the click pairs for gaps down to 2–3 ms. Thus, the limit to temporal resolution found in this task is similar to that found for the detection of a gap in broadband noise. It should be noted that, in this task, subjects do not hear the individual clicks within a click pair. Rather, each click pair is heard as a single sound with its own characteristic quality.

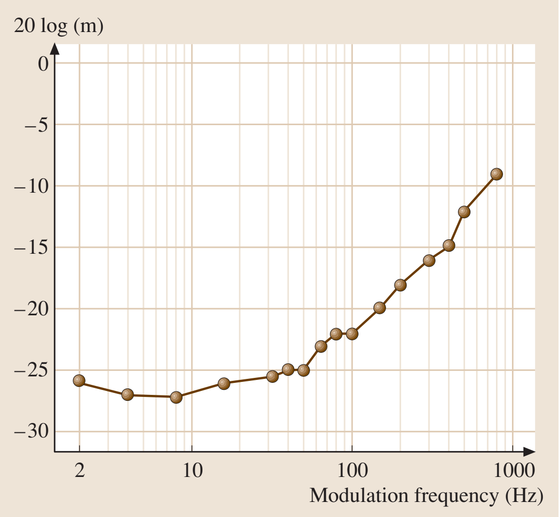

Fig. 13.15 A temporal modulation transfer function (TMTF). A broadband white noise was sinusoidally amplitude modulated, and the threshold amount of modulation required for detection is plotted as a function of modulation rate. The amount of modulation is specified as $20\log(m)$, where m is the modulation index (see Chap. 14). The higher the sensitivity to modulation, the more negative is this quantity. (After *Bacon* and *Viemeister* [13.70])

The experiments described above each give a single number to describe temporal resolution. A more comprehensive approach is to measure the threshold for detecting changes in the amplitude of a sound as a function of the rapidity of the changes. In the simplest case, white noise is sinusoidally amplitude modulated, and the threshold for detecting the modulation is determined as a function of modulation rate. The function relating threshold to modulation rate is known as a temporal modulation transfer function (TMTF). Modulation of white noise does not change its long-term magnitude spectrum. An example of the results is shown in Fig. 13.15; data are taken from *Bacon* and *Viemeister* [13.70]. For low modulation rates, performance is limited by the amplitude resolution of the ear, rather than by temporal resolution. Thus, the threshold is independent of modulation rate for rates up to about 50 Hz. As the rate increases beyond 50 Hz, temporal resolution starts to have an effect; the threshold increases, and for rates above about 1000 Hz the modulation cannot be detected at all. Thus, sensitivity to amplitude modulation decreases progressively as the rate of modulation increases. The shapes of TMTFs do not vary much with overall sound level, but the ability to detect the modulation does worsen at low sound levels.

To explore whether temporal resolution varies with center frequency, *Green* [13.75] used stimuli which consisted of a brief pulse of a sinusoid in which the level of the first half of the pulse was 10 dB different from that of the second half. Subjects were required to distinguish two signals, differing in whether the half with the high level was first or second. Green measured performance as a function of the total duration of the stimuli. The threshold, corresponding to 75% correct discrimination, was similar for center frequencies of 2 and 4 kHz and was between 1 and 2 ms. However, the threshold was slightly higher for a center frequency of 1 kHz, being between 2 and 4 ms.

Moore et al. [13.76] measured the threshold for detecting a gap in a sinusoid, for signal frequencies of 100, 200, 400, 800, 1000, and 2000 Hz. A background noise was used to mask the spectral *splatter* produced by turning the sound off and on to produce the gap. The gap thresholds were almost constant, at 6–8 ms over the frequency range 400–2000 Hz, but increased somewhat at 200 Hz and increased markedly, to about 18 ms, at 100 Hz. Individual variability also increased markedly at 100 Hz.

Overall, it seems that temporal resolution is roughly independent of frequency for medium to high frequencies, but worsens somewhat at very low center frequencies.

The measurement of TMTFs using sinusoidal carriers is complicated by the fact that the modulation introduces spectral sidebands, which may be detected as separate components if they are sufficiently far in frequency from the carrier frequency. When the carrier frequency is high, the effect of resolution of sidebands is likely to be small for modulation frequencies up to a few hundred Hertz, as the auditory filter bandwidth is large for high center frequencies. Consistent with this, TMTFs for high carrier frequencies generally show an initial flat portion (sensitivity independent of modulation frequency), then a portion where threshold increases with increasing modulation frequency, presumably reflecting the limits of temporal resolution, and then a portion where threshold decreases again, presumably reflecting the detection of spectral sidebands [13.77, 78].

The initial flat portion of the TMTF extends to about 100–120 Hz for sinusoidal carriers, but only to 50 or 60 Hz for a broadband noise carrier. It has been suggested that the discrepancy occurs because the inherent amplitude fluctuations in a noise carrier limit the detectability of the imposed modulation [13.79–82]; see below for further discussion of this point. The effect of the inherent fluctuations depends upon their similarity to the imposed modulation. When a narrow-band noise carrier is used, which has relatively slow inherent amplitude fluctuations, TMTFs show the poorest sensitivity for low modulation frequencies [13.79, 80]. In principle, then, TMTFs obtained using sinusoidal carriers provide a better measure of the inherent temporal resolution of the auditory system than TMTFs obtained using noise carriers, provided that the modulation frequency is within the range where spectral resolution does not play a major role.

13.4.2 Modeling Temporal Resolution

Most models of temporal resolution are based on the idea that there is a process at levels of the auditory system higher than the auditory nerve which is sluggish in some way, thereby limiting temporal resolution. The models assume that the internal representation of stimuli is smoothed over time, so that rapid temporal changes are reduced in magnitude but slower ones are preserved. Although this smoothing process almost certainly operates on neural activity, the most widely used models are based on smoothing a simple transformation of the stimulus, rather than its neural representation.

Most models include an initial stage of bandpass filtering, reflecting the action of the auditory filters. Each filter is followed by a nonlinear device. This nonlinear device is meant to reflect the operation of several processes that occur in the peripheral auditory system such as compression on the basilar membrane and neural transduction, whose effects resemble half-wave rectification (see Chap. 12). The output of the nonlinear device is fed to a smoothing device, which can be implemented either as a low-pass filter [13.83] or (equivalently) as a sliding temporal integrator [13.38, 84]. The device determines a kind of weighted average of the output of the compressive nonlinearity over a certain time interval or window. This weighting function is sometimes called the shape of the temporal window. The window itself is assumed to slide in time, so that the output of the temporal integrator is like a weighted running average of the input. This has the effect of smoothing rapid fluctuations while preserving slower ones. When a sound is turned on abruptly, the output of the temporal integrator takes some time to build up. Similarly, when a sound is turned off, the output of the integrator takes some time to decay. The shape of the window is assumed to be asymmetric in time, such that the build up of its output in response to the onset of a sound is more rapid than the decay of its output in response to the cessation of a sound. The output of the sliding temporal integrator is fed to a decision device. The decision device may use different rules depending on the task required. For example, if the task is to detect a brief temporal gap in a signal, the decision device might look for a dip in the output of the temporal integrator. If the task is to detect amplitude modulation of a sound, the device might assess the amount of modulation at the output of the sliding temporal integrator [13.83].

It is often assumed that backward and forward masking depend on the process of build up and decay. For example, if a brief signal is rapidly followed by a masker (backward masking), the response to the signal may still be building up when the masker occurs. If the masker is sufficiently intense, then its internal effects may swamp those of the signal. Similarly, if a brief signal follows soon after a masker (forward masking), the decaying response to the masker may swamp the response to the signal. The asymmetry in the shape of the window accounts for the fact that forward masking occurs over longer masker-signal intervals than backward masking.

13.4.3 A Modulation Filter Bank?

Some researchers have suggested that the analysis of sounds that are amplitude modulated depends on a specialized part of the brain that contains an array of neurons, each tuned to a different modulation rate [13.85]. Each neuron can be considered as a filter in the modulation domain, and the array of neurons is known collectively as a modulation filter bank. The modulation filter bank has been suggested as a possible explanation for certain perceptual phenomena, which are described below. It should be emphasized, however, that this is still a controversial concept.

The threshold for detecting amplitude modulation of a given carrier generally increases if additional amplitude modulation is superimposed on that carrier. This effect has been called modulation masking. *Houtgast* [13.86] studied the detection of sinusoidal amplitude modulation of a pink noise carrier. Thresholds for detecting the modulation were measured when no other modulation was present and when a masker modulator was present in addition. In one experiment, the masker modulation was a half-octave-wide band of noise, with a center frequency of 4, 8, or 16 Hz. For each masker, the masking pattern showed a peak at the masker frequency. This could be interpreted as indicating selectivity in the modulation-frequency domain, analogous to the frequency selectivity in the audio-frequency domain that was described earlier.

Bacon and *Grantham* [13.87] measured thresholds for detecting sinusoidal amplitude modulation of a broadband white noise in the presence of a sinusoidal masker modulator. When the masker modulation frequency was 16 or 64 Hz, most modulation masking occurred when the signal modulation frequency was near the masker frequency. In other words, the masking patterns were roughly bandpass, although they showed an increase for very low signal frequencies. For a 4 Hz masker, the masking patterns had a low-pass characteristic, i. e., there was a downward spread of modulation masking.

It should be noted that the sharpness of tuning of the hypothetical modulation filter bank is much less than the sharpness of tuning of the auditory filters in the audio-frequency domain. The bandwidths have been estimated as between 0.5 and 1 times the center frequency [13.80, 88–90]. The modulation filters, if they exist, are not highly selective.

13.4.4 Duration Discrimination

Duration discrimination has typically been studied by presenting two successive sounds which have the same power spectrum but differ in duration. The subject is required to indicate which sound had the longer duration. Both *Creelman* [13.91] and *Abel* [13.92] found that the smallest detectable increase in duration, ΔT, increased with the baseline duration T. The data of Abel showed that, for $T = 10$, 100, and 1000 ms, ΔT was about 4, 15, and 60 ms, respectively. Thus, the Weber fraction, $\Delta T/T$, decreased with increasing T. The results were relatively independent of the overall level of the stimuli and were similar for noise bursts of various bandwidths and for bursts of a 1000 Hz sine wave.

Abel [13.93] reported somewhat different results for the discrimination of the duration of the silent interval between two *markers*. For silent durations, T, less than 160 ms, the results showed that discrimination improved as the level of the markers increased. The function relating $\Delta T/T$ to T was non-monotonic, reaching a local minimum for $T = 2.5$ ms and a local maximum for $T = 10$ ms. The value of ΔT ranged from 6 to 19 ms for a base duration of 10 ms and from 61 to 96 ms for a base duration of 320 ms.

Divenyi and *Danner* [13.94] required subjects to discriminate the duration of the silent interval defined by two 20 ms sounds. When the markers were identical high-level (86 dB SPL) bursts of sinusoids or noise, performance was similar across markers varying in center frequency (500–4000 Hz) and bandwidth. In contrast to the results of *Abel* [13.93], $\Delta T/T$ was almost independent of T over the range of T from 25 to 320 ms. Thresholds were markedly lower than those reported by *Abel* [13.93], being about 1.7 ms at $T = 25$ ms and 15 ms at $T = 320$ ms. This may have been a result of the extensive training of the subjects of Divenyi and Danner. For bursts of a 1 kHz sinusoid, performance worsened markedly when the level of the markers was decreased below 25 dB SL. Performance also worsened markedly when the two markers on either side of a silent interval were made different in level or frequency.

In summary, all studies show that, for values of T exceeding 10 ms, ΔT increases with T and ΔT is roughly independent of the spectral characteristics of the sounds. This is true both for the duration discrimination of sounds and for the discrimination of silent intervals bounded by acoustic markers, provided that the markers are identical on either side of the interval. However, ΔT increases at low sound levels, and also increases when the markers differ in level or frequency on either side of the interval.

13.4.5 Temporal Analysis Based on Across-Channel Processes

Studies of the ability to compare timing across different frequency channels can give very different results depending on whether the different frequency components in the sound are perceived as part of a single sound or as part of more than one sound. Also, it should be realized that subjects may be able to *distinguish* different time patterns, for example, a change in the relative onset time of two different frequencies, without the subjective impression of a change in time pattern; some sort of change in the quality of the sound may be all that is heard. The studies described next indicate the limits of the ability to compare timing across channels, using highly trained subjects.

Patterson and *Green* [13.95] and *Green* [13.75] have studied the discrimination of a class of signals which have the same long-term magnitude spectrum, but which differ in their short-term spectra. These sounds, called Huffman sequences, are brief, broadband click-like sounds, except that the energy in a certain frequency region is delayed relative to that in other regions. The amount of the delay, the center frequency of the delayed frequency region, and the width of the delayed frequency region can all be varied. If subjects can distinguish a pair of Huffman sequences differing, for example, in the amount of delay in a given frequency region, this implies that they are sensitive to the difference in time pattern, i. e., they must be detecting that one frequency region is delayed relative to other regions. *Green* [13.75] measured the ability of subjects to detect differences in the amount of delay in three frequency regions: 650, 1900 and 4200 Hz. He found similar results for all three center frequencies: subjects could detect differences in delay time of about 2 ms regardless of the center frequency of the delayed region.

It should be noted that subjects did not report hearing one part of the sound after the rest of the sound. Rather, the differences in time pattern were perceived as subtle changes in sound quality. Further, some subjects required extensive training to achieve the fine acuity of 2 ms, and even after this training the task required considerable concentration.

Zera and *Green* [13.96] measured thresholds for detecting asynchrony in the onset or offset of complex signals composed of many sinusoidal components.

The components were either uniformly spaced on a logarithmic frequency scale or formed a harmonic series. In one stimulus, the standard, all components started and stopped synchronously. In the signal stimulus, one component was presented with an onset or offset asynchrony. The task of the subjects was to discriminate the standard stimulus from the signal stimulus. They found that onset asynchrony was easier to detect than offset asynchrony. For harmonic signals, onset asynchronies less than 1 ms could generally be detected, whether the asynchronous component was leading or lagging the other components. Thresholds for detecting offset asynchronies were larger, being about 3–10 ms when the asynchronous component ended after the other components and 10–30 ms when the asynchronous component ended before the other components. Thresholds for detecting asynchronies in logarithmically spaced complexes were generally 2–50 times larger than for harmonic complexes.

The difference between harmonically and logarithmically spaced complexes may be explicable in terms of perceptual grouping (see later for more details). The harmonic signal was perceived as a single sound source, i. e., all of the components appeared to belong together. The logarithmically spaced complex was perceived as a series of separate tones, like many notes being played at once on an organ. It seems that it is difficult to compare the timing of sound elements that are perceived as coming from different sources, a point that will be expanded later. The high sensitivity to onset asynchronies for harmonic complexes is consistent with the finding that the perceived timbres of musical instruments are partly dependent on the exact onset times and rates of rise of individual harmonics within each musical note [13.97]. This is described in more detail later on.

13.5 Pitch Perception

Pitch is an attribute of sound defined in terms of what is *heard*. It is defined formally as “that attribute of auditory sensation in terms of which sounds can be ordered on a scale extending from low to high” [13.10]. It is related to the physical repetition rate of the waveform of a sound; for a pure tone (a sinusoid) this corresponds to the frequency, and for a periodic complex tone to the fundamental frequency. Increasing the repetition rate gives a sensation of increasing pitch. Appropriate variations in repetition rate can give rise to a sense of melody. Variations in pitch are also associated with the intonation of voices, and they provide cues as to whether an utterance is a question or a statement and as to the emotion of the talker. Since pitch is a subjective attribute, it cannot be measured directly. Often, the pitch of a complex sound is assessed by adjusting the frequency of a sinusoid until the pitch of the sinusoid matches the pitch of the sound in question. The frequency of the sinusoid then gives a measure of the pitch of the sound. Sometimes a periodic complex sound, such as a pulse train, is used as a matching stimulus. In this case, the repetition rate of the pulse train gives a measure of pitch. Results are generally similar for the two methods, although it is easier to make a pitch match when the sounds to be matched do not differ very much in timbre (see Sect. 13.6.)

13.5.1 Theories of Pitch Perception

There are two traditional theories of pitch perception. One, the *place* theory, is based on the fact that different frequencies (or frequency components in a complex sound) excite different places along the basilar membrane, and hence neurons with different CFs. The place theory assumes that the pitch of a sound is related to the excitation pattern produced by that sound; for a pure tone the pitch is generally assumed to be determined by the position of maximum excitation [13.98].

An alternative theory, called the *temporal* theory, is based on the assumption that the pitch of a sound is related to the time pattern of the neural impulses evoked by that sound [13.99]. These impulses tend to occur at a particular phase of the waveform on the basilar membrane, a phenomenon called phase locking (see Chap. 18). The intervals between successive neural impulses approximate integer multiples of the period of the waveform and these intervals are assumed to determine the perceived pitch. The temporal theory cannot be applicable at very high frequencies, since phase locking does not occur for frequencies above about 5 kHz. However, the tones produced by most musical instruments, the human voice, and most everyday sound sources have fundamental frequencies well below this range.

Many researchers believe that the perception of pitch involves both place mechanisms and temporal mechanisms. However, one mechanism may be dominant for a specific task or aspect of pitch perception, and the relative role of the two mechanisms almost certainly varies with center frequency.

13.5.2 The Perception of the Pitch of Pure Tones

The Frequency Discrimination of Pure Tones

It is important to distinguish between frequency selectivity and frequency discrimination. The former refers to the ability to resolve the frequency components of a complex sound. The latter refers to the ability to detect changes in frequency over time. Usually, the changes in frequency are heard as changes in pitch. The smallest detectable change in frequency is called the frequency difference limen (DL).

Place models of frequency discrimination [13.100, 101] predict that frequency discrimination should be related to frequency selectivity; both should depend on the sharpness of tuning on the basilar membrane. *Zwicker* [13.101] described a model of frequency discrimination based on changes in the excitation pattern evoked by the sound when the frequency is altered, inferring the shapes of the excitation patterns from masking patterns such as those shown in Fig. 13.6. The model is illustrated in Fig. 13.16. The figure shows two excitation patterns, corresponding to two tones with slightly different frequencies. A change in frequency results in a sideways shift of the excitation pattern. The change is assumed to be detectable whenever the excitation level at some point on the excitation pattern changes by more than about 1 dB.

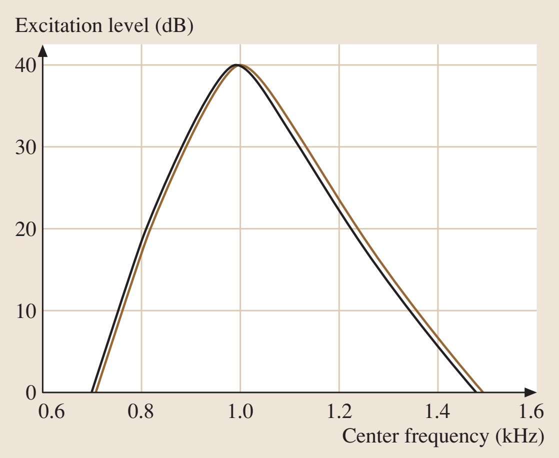

Fig. 13.16 Schematic illustration of an excitation-pattern model for frequency discrimination. Excitation patterns are shown for two sinusoidal tones differing slightly in frequency; the two tones have frequencies of 995 Hz and 1005 Hz. It is assumed that the difference in frequency, ΔF, can be detected if the excitation level changes anywhere by more than a criterion amount. The biggest change in excitation level is on the low-frequency side

The change in excitation level is greatest on the steeply sloping low-frequency (low-CF) side of the excitation pattern. Thus, in this model, the detection of a change in frequency is functionally equivalent to the detection of a change in level on the low-frequency side of the excitation pattern. The steepness of the low-frequency side is roughly constant when the frequency scale is expressed in units of $\mathrm{ERB_N}$, rather than in terms of linear frequency. The slope is about 18 dB per $\mathrm{ERB_N}$. To achieve a change in excitation level of 1 dB, the frequency has to be changed by one eighteenth of an $\mathrm{ERB_N}$. Thus, Zwicker's model predicts that the frequency DL at any given frequency should be about one eighteenth (= 0.056) of the value of $\mathrm{ERB_N}$ at that frequency.

To test Zwicker's model, frequency DLs have been measured as a function of center frequency. There have been two common ways of measuring frequency discrimination. One involves the discrimination of two successive steady tones with slightly different frequencies. On each trial, the tones are presented in a random order and the listener is required to indicate whether the first or second tone is higher in frequency. The frequency difference between the two tones is adjusted until the listener achieves a criterion percentage correct, for example 75%. This measure will be called the difference limen for frequency (DLF). A second measure, called the frequency modulation detection limen (FMDL), uses tones which are frequency modulated. In such tones, the frequency moves up and down in a regular periodic manner about the mean (carrier) frequency. The number of times per second that the frequency goes up and down is called the modulation rate. Typically, the modulation rate is rather low (2–20 Hz), and the change in frequency is heard as a fluctuation in pitch – a kind of *warble*. To determine a threshold for detecting frequency modulation, two tones are presented successively; one is modulated in frequency and the other has a steady frequency. The order of the tones on each trial is random. The listener is required to indicate whether the first or the second tone is modulated. The amount of modulation (also called the modulation depth) required to achieve a criterion response (e.g. 75% correct) is determined.

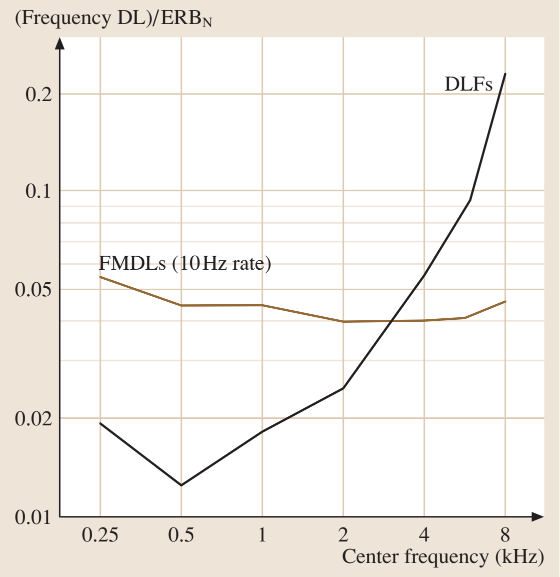

Fig. 13.17 Thresholds for detecting differences in frequency between steady pulsed tones (DLFs) and for detecting frequency modulation (FMDLs), plotted relative to the ERB$_N$ of the auditory filter at each center frequency. (After *Sek* and *Moore* [13.103])

It turns out that the results obtained with these two methods are quite different [13.102], although the difference depends on the modulation rate used to measure the FMDLs. An example of results obtained with the two methods is given in Fig. 13.17 (data from *Sek* and *Moore* [13.103]). For the FMDLs the modulation rate was 10 Hz. To test Zwicker's model, the DLFs and FMDLs were plotted as a proportion of the value of ERB$_N$ at the same center frequency. According to this model, the proportion should be independent of frequency. The proportion for FMDLs using a 10 Hz modulation rate, shown as the brighter line, is roughly constant, and its value is about 0.05, close to the value predicted by the model. However, the proportion for DLFs varies markedly with frequency [13.103, 104]. This is illustrated by the dark line in Fig. 13.17. The DLFs for frequencies of 2 kHz and below are smaller than predicted by Zwicker's model, while those for frequencies of 6 and 8 kHz are larger than predicted.

The results for the FMDLs are consistent with the place model, but the results for the DLFs are not. The reason for the deviation for DLFs is probably that DLFs at low frequencies depend on the use of temporal information from phase locking. Phase locking becomes less precise at frequencies above 1 kHz, and it is completely lost above 5 kHz. This can account for the marked increase in the DLFs at high frequencies [13.105].

The ratio FMDL/ERB$_N$ is not constant across center frequency when the modulation rate is very low (around 2 Hz), but increases with increasing frequency [13.103, 106]. For low center frequencies, FMDLs are smaller for a 2 Hz modulation rate than for a 10 Hz rate, while for high carrier frequencies (above 4 kHz) the reverse is true. Thus, for a 2 Hz modulation rate, the agreement between DLFs and FMDLs is better than for a 10 Hz rate, but discrepancies remain. For very low modulation rates, frequency modulation may be detected by virtue of the changes in phase locking to the carrier that occur over time. In other words, the frequency is determined over short intervals of time, using phase-locking information, and changes in the estimated frequency over time indicate the presence of frequency modulation. *Moore* and *Sek* [13.106] suggested that the mechanism for decoding the phase-locking information was sluggish; it had to sample the sound for a certain time in order to estimate its frequency. Hence, it could not follow rapid changes in frequency and it played little role for high modulation rates.

In summary, measures of frequency discrimination are consistent with the idea that DLFs, and FMDLs for very low modulation rates, are determined by temporal information (phase locking) for frequencies up to about 4–5 kHz. The precision of phase locking decreases with increasing frequency above 1–2 kHz, and it is almost absent above about 5 kHz. This can explain why DLFs increase markedly at high frequencies. FMDLs for medium to high modulation rates may be determined by a place mechanism, i. e., by the detection of changes in the excitation pattern. This mechanism may also account for DLFs and for FMDLs for low modulation rates, when the center frequency is above about 5 kHz.

The Perception of Musical Intervals

If temporal information plays a role in determining the pitch of pure tones, then we would expect changes in perception to occur for frequencies above 5 kHz, at which phase locking does not occur. Two aspects of perception do indeed change in the expected way, namely the perception of musical intervals, and the perception of melodies.

Two tones which are separated in frequency by an interval of one *octave* (i. e., one has twice the frequency of the other) sound similar. They are judged to have the same name on the musical scale (for example, C3 and C4). This has led several theorists to suggest that there are at least two dimensions to musical pitch. One aspect is related monotonically to frequency (for a pure tone) and is known as *tone height*. The other is related to pitch class (i. e., the name of the note) and is called *tone chroma* [13.107, 108]. For example, two sinusoids with frequencies of 220 and 440 Hz would have the same tone chroma (they would both be called A on the musical scale) but, as they are separated by an octave, they would have different tone heights.

If subjects are presented with a pure tone of a given frequency, f_1, and are asked to adjust the frequency, f_2 of a second tone (presented so as to alternate in time with the fixed tone) so that it appears to be an octave higher in pitch, they generally adjust f_2 to be roughly twice f_1. However, when f_1 lies above 2.5 kHz, so that f_2 would lie above 5 kHz, octave matches become very erratic [13.109]. It appears that the musical interval of an octave is only clearly perceived when both tones are below 5 kHz.

Other aspects of the perception of pitch also change above 5 kHz. A sequence of pure tones above 5 kHz does not produce a clear sense of melody [13.110]. It is possible to hear that the pitch changes when the frequency is changed, but the musical intervals are not heard clearly. Also, subjects with absolute pitch (the ability to assign names to notes without reference to other notes) are very poor at naming notes above 4–5 kHz [13.111].

These results are consistent with the idea that the pitch of pure tones is determined by different mechanisms above and below 5 kHz, specifically, by a temporal mechanism at low frequencies and a place mechanism at high frequencies. It appears that the perceptual dimension of tone height persists over the whole audible frequency range, but tone chroma only occurs in the frequency range below 5 kHz. Musical intervals are only clearly perceived when the frequencies of the tones lie in the range where temporal information is available.

The Effect of Level on Pitch

The pitch of a pure tone is primarily determined by its frequency. However, sound level also plays a small role. On average, the pitch of tones with frequencies below about 2 kHz decreases with increasing level, while the pitch of tones with frequencies above about 4 kHz increases with increasing sound level. The early data of *Stevens* [13.112] showed rather large effects of sound level on pitch, but other data generally show much smaller effects [13.113]. For tones with frequencies between 1 and 2 kHz, changes in pitch with level are generally less than 1%. For tones of lower and higher frequencies, the changes can be larger (up to 5%). There are also considerable individual differences both in the size of the pitch shifts with level, and in the direction of the shifts [13.113].

13.5.3 The Perception of the Pitch of Complex Tones

The Phenomenon of the Missing Fundamental

For complex tones the pitch does not, in general, correspond to the position of maximum excitation on the basilar membrane. Consider, as an example, a sound consisting of short impulses (clicks) occurring 200 times per second. This sound contains harmonics with frequencies at integer multiples of 200 Hz (200, 400, 600, 800 ... Hz). The harmonic at 200 Hz is called the fundamental frequency. The sound has a low pitch, which is very close to the pitch of its fundamental component (200 Hz), and a sharp timbre (a buzzy tone quality). However, if the sound is filtered so as to remove the fundamental component, the pitch does not alter; the only result is a slight change in timbre. This is called the phenomenon of the missing fundamental [13.99, 114]. Indeed, all except a small group of mid-frequency harmonics can be eliminated, and the low pitch remains the same, although the timbre becomes markedly different.

Schouten [13.99,115] called the low pitch associated with a group of high harmonics the *residue*. Several other names have been used to describe this pitch, including *periodicity pitch*, *virtual pitch*, and *low pitch*. The term *low pitch* will be used here. Schouten pointed out that it is possible to hear the change produced by removing the fundamental component and then reintroducing it. Indeed, when the fundamental component is present, it is possible to hear it out as a separate sound. The pitch of that component is almost the same as the pitch of the whole sound. Therefore, the presence or absence of the fundamental component does not markedly affect the pitch of the whole sound.

The perception of a low pitch does not require activity at the point on the basilar membrane which would respond maximally to the fundamental component. *Licklider* [13.116] showed that the low pitch could be heard when low-frequency noise was present that would mask any component at the fundamental frequency. Even when the fundamental component of a complex tone is present, the pitch of the tone is usually determined by harmonics other than the fundamental.

The phenomenon of the missing fundamental is not consistent with a simple place model of pitch based on the idea that pitch is determined by the position of the peak excitation on the basilar membrane. However, more elaborate place models have been proposed, and these are discussed below.

Theories of Pitch Perception for Complex Tones

To understand theories of pitch perception for complex tones, it is helpful to consider how complex tones are represented in the peripheral auditory system. A simulation of the response of the basilar membrane to a complex tone is illustrated in Fig. 13.18. In this example, the complex tone is a regular series of brief pulses, whose spectrum contains many equal-amplitude harmonics. The number of pulses per second (also called the repetition rate) is 200, so the harmonics have frequencies that are integer multiples of 200 Hz. The lower harmonics are partly resolved on the basilar membrane, and give rise to distinct peaks in the pattern of activity along the basilar membrane. At a place tuned to the frequency of a low harmonic, the waveform on the basilar membrane is approximately a sinusoid at the harmonic frequency. For example, at the place with a characteristic frequency of 400 Hz the waveform is a 400 Hz sinusoid. At a place tuned between two low harmonics, e.g., the place tuned to 317 Hz, there is very little response. In contrast, the higher harmonics are not resolved, and do not give rise to distinct peaks on the basilar membrane. The waveforms at places on the basilar membrane responding to higher harmonics are complex, but they all have a repetition rate equal to the fundamental frequency of the sound.

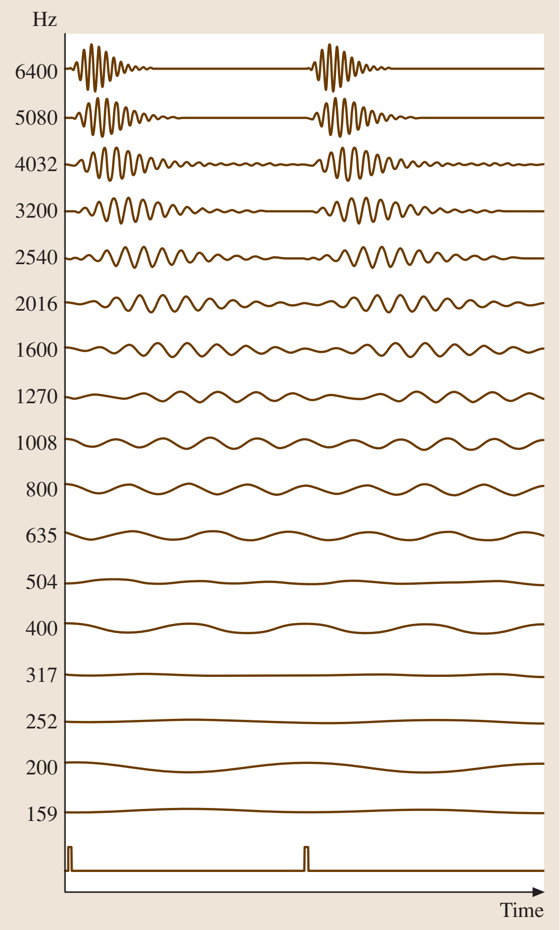

Fig. 13.18 A simulation of the responses on the basilar membrane to periodic impulses of rate 200 pulses per second. The input waveform is shown at the bottom; impulses occur every 5 ms. Each number on the left represents the frequency which would maximally excite a given point on the basilar membrane. The waveform which would be observed at that point, as a function of time, is plotted opposite that number

There are two main (nonexclusive) ways in which the low pitch of a complex sound might be extracted. Firstly, it might be derived from the frequencies of the lower harmonics that are resolved on the basilar membrane. The frequencies of the harmonics might be determined either by place mechanisms (e.g., from the positions of local maxima on the basilar membrane) or by temporal mechanisms (from the inter-spike intervals in neurons with CFs close to the frequencies of individual harmonics). For example, for the complex tone whose analysis is illustrated in Fig. 13.18, the second harmonic, with a frequency of 400 Hz, would give rise to a local maximum at the place on the basilar membrane tuned to 400 Hz. The inter-spike intervals in neurons innervating that place would reflect the frequency of that harmonic; the intervals would cluster around integer multiples of 2.5 ms. Both of these forms of information may allow the auditory system to determine that there is a harmonic at 400 Hz.

The auditory system may contain a pattern recognizer which determines the low pitch of the complex sound from the frequencies of the resolved components [13.117–119]. In essence the pattern recognizer tries to find the harmonic series giving the best match to the resolved frequency components; the fundamental frequency of this harmonic series determines the perceived pitch. Say, for example, that the initial analysis establishes frequencies of 800, 1000 and 1200 Hz to be present. The fundamental frequency whose harmonics would match these frequencies is 200 Hz. The perceived pitch corresponds to this inferred fundamental frequency of 200 Hz. Note that the inferred fundamental frequency is always the highest possible value that fits the frequencies determined in the initial analysis. For example, a fundamental frequency of 100 Hz would also have harmonics at 800, 1000 and 1200 Hz, but a pitch corresponding to 100 Hz is *not* perceived. It should also be noted that when a complex tone contains only two or three low harmonics, some people, called *analytic listeners*, do not hear the low pitch, but rather hear pitches corresponding to individual harmonics [13.120, 121]. Others, called *synthetic listeners*, usually hear only the low pitch. When many harmonics are present, the low pitch is usually the dominant percept.

Evidence supporting the idea that the low pitch of a complex tone is derived by combining information from several harmonics comes from studies of the ability to detect changes in repetition rate (equivalent to the number of periods per second). When the repetition rate of a complex tone changes, all of the components change in frequency by the same *ratio*, and a change in low pitch is heard. The ability to detect such changes is better than the ability to detect changes in a sinusoid at the fundamental frequency [13.122] and it can be better than the ability to detect changes in the frequency of any of the individual sinusoidal components in the complex tone [13.123]. This indicates that information from the different harmonics is combined or integrated in the determination of low pitch. This can lead to very fine discrimination; changes in repetition rate of about 0.2% can be detected for fundamental frequencies in the range 100–400 Hz provided that low harmonics are present (e.g., the third, fourth and fifth).

The pitch of a complex tone may also be extracted from the higher unresolved harmonics. As shown in Fig. 13.18, the waveforms at places on the basilar membrane responding to higher harmonics are complex, but they all have a repetition rate equal to the fundamental frequency of the sound, namely 200 Hz. For the neurons with CFs corresponding to the higher harmonics, nerve impulses tend to be evoked by the biggest peaks in the waveform, i. e., by the waveform peaks close to envelope maxima. Hence, the nerve impulses are separated by times corresponding to the period of the sound. For example, in Fig. 13.18 the input has a repetition rate of 200 periods per second, so the period is 5 ms. The time intervals between nerve spike would cluster around integer multiples of 5 ms, i. e., 5, 10 15, 20 ... ms. The pitch may be determined from these time intervals. In this example, the time intervals are integer multiples of 5 ms, so the pitch corresponds to 200 Hz.

Experimental evidence suggests that pitch can be extracted *both* from the lower harmonics and from the higher harmonics. Usually, the lower, resolved harmonics give a clearer low pitch, and are more important in determining low pitch, than the upper unresolved harmonics [13.124–126]. This idea is called the principle of dominance; when a complex tone contains many harmonics, including both low- and high-numbered harmonics, the pitch is mainly determined by a small group of lower harmonics. Also, the discrimination of changes in repetition rate of complex tones is better for tones containing only low harmonics than for tones containing only high harmonics [13.123, 127–129]. However, a low pitch can be heard when only high unresolvable harmonics are present. Although, this pitch is not as clear as when lower harmonics are present, it is clear enough to allow the recognition of musical intervals and of simple melodies [13.129, 130].

Several researchers have proposed theories in which both place (spectral) and temporal mechanisms play a role; these are referred to as spectro-temporal theories. The theories assume that information from both low harmonics and high harmonics contributes to the determination of pitch. The initial place/spectral analysis in the cochlea is followed by an analysis of the time pattern of the neural spikes evoked at each CF [13.131–135]. The temporal analysis is assumed to occur at a level of the auditory system higher than the auditory nerve, perhaps in the cochlear nucleus. In the model proposed by *Moore* [13.133], the sound is passed through an array of bandpass filters, each corresponding to a specific place on the basilar membrane. The time pattern of the neural impulses at each CF is determined by the waveform at the corresponding point on the basilar membrane. The inter-spike intervals at each CF are determined. Then, a device compares the time intervals present at different CFs, and searches for common time intervals. The device may also integrate information over time. In general the time interval which is found most often corresponds to the period of the fundamental component. The perceived pitch corresponds to the reciprocal of this interval. For example, if the most prominent time interval is 5 ms, the perceived pitch corresponds to a frequency of 200 Hz.

13.6 Timbre Perception

13.6.1 Time-Invariant Patterns and Timbre

Almost all of the sounds that we encounter in everyday life contain a multitude of frequency components with particular levels and relative phases. The distribution of energy over frequency is one of the major determinants of the quality of a sound or its timbre. Timbre is usually defined as "that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar" [13.10]. Differences in timbre enable us to distinguish between the same note played on, say, the piano, the violin or the flute.

Timbre depends upon more than just the frequency spectrum of the sound; fluctuations over time can play an important role (see the next section). For the purpose of this section we can adopt a more restricted definition suggested by *Plomp* [13.136] as applicable to steady tones: "Timbre is that attribute of sensation in terms of which a listener can judge that two steady complex tones having the same loudness, pitch and duration are dissimilar." Timbre defined in this way depends mainly on the relative magnitudes of the partials of the tones.

Timbre is multidimensional; there is no single scale along which the timbres of different sounds can be compared or ordered. Thus, a way is needed of describing the spectrum of a sound which takes into account this multidimensional aspect, and which can be related to the subjective timbre. A crude first approach is to look at the overall distribution of spectral energy. The *brightness* or *sharpness* [13.137] of sounds seems to be related to the spectral centroid. However, a much more quantitative approach has been described by *Plomp* and his colleagues [13.136, 138]. They showed that the perceptual differences between different sounds, such as vowels, or steady tones produced by musical instruments, were closely related to the differences in the spectra of the sounds, when the spectra were specified as the levels in 18 1/3-octave frequency bands. A bandwidth of 1/3 octave is slightly greater than the ERB_N of the auditory filter over most of the audible frequency range. Thus, timbre is related to the relative level produced at the output of each auditory filter. Put another way, the timbre of a sound is related to the excitation pattern of that sound.

It is likely that the number of dimensions required to characterize timbre is limited by the number of ERB_Ns required to cover the audible frequency range. This would give a maximum of about 37 dimensions. For a restricted class of sounds, however, a much smaller number of dimensions may be involved. It appears to be generally true, both for speech and non-speech sounds, that the timbres of steady tones are determined primarily by their magnitude spectra, although the relative phases of the components may also play a small role [13.139, 140].

13.6.2 Time-Varying Patterns and Auditory Object Identification

Differences in static timbre are not always sufficient to allow the absolute identification of an *auditory object*, such as a musical instrument. One reason for this is that the magnitude and phase spectrum of the sound may be markedly altered by the transmission path and room reflections [13.141]. In practice, the recognition of a particular timbre, and hence of an auditory object, may depend upon several other factors. *Schouten* [13.142] has suggested that these include: (1) whether the sound is periodic, having a tonal quality for repetition rates between about 20 and 20 000 per second, or irregular, and having a noise-like character; (2) whether the waveform envelope is constant, or fluctuates as a function of time, and in the latter case what the fluctuations are like; (3) whether any other aspect of the sound (e.g., spectrum or periodicity) is changing as a function of time; (4) what the preceding and following sounds are like.

The recognition of musical instruments, for example, depends quite strongly on onset transients and on the temporal structure of the sound envelope. The characteristic tone of a piano depends upon the fact that the notes have a rapid onset and a gradual decay. If a recording of a piano is reversed in time, the timbre is completely different. It now resembles that of a harmonium or accordion, in spite of the fact that the long-term magnitude spectrum is unchanged by time reversal. The perception of sounds with temporally asymmetric envelopes has been studied by *Patterson* [13.143, 144]. He used sinusoidal carriers that were amplitude modulated by a repeating exponential function. The envelope either increased abruptly and decayed gradually (damped sounds) or increased gradually and decayed abruptly (ramped sounds). The ramped sounds were time-reversed versions of the damped sounds and had the same long-term magnitude spectrum. The sounds were characterized by the repetition period of the envelope, which was 25 ms, and by the half life. For a damped sinusoid, the half life is the time taken for the amplitude to decrease by a factor of two.

Patterson reported that the ramped and damped sounds had different qualities. For a half life of 4 ms, the damped sound was perceived as a single source rather like a drum roll played on a hollow, resonant surface (like a drummer's wood block). The ramped sound was perceived as two sounds: a drum roll on a non-resonant surface (such as a leather table top) and a continuous tone corresponding to the carrier frequency. *Akeroyd* and *Patterson* [13.145] used sounds with similar envelopes, but the carrier was broadband noise rather than a sinusoid. They reported that the damped sound was heard as a drum struck by wire brushes. It did not have any hiss-like quality. In contrast, the ramped sound was heard as a noise, with a hiss-like quality, that was sharply cut off in time. These experiments clearly demonstrate the important role of temporal envelope in timbre perception

Pollard and *Jansson* [13.146] described a perceptually relevant way of characterizing the characteristics of time-varying sounds. The sound is filtered in bands of width 1/3 octave. The loudness of each band is calculated at 5 ms intervals. The loudness values are then converted into three coordinates, based on the loudness of:

1. the fundamental component,
2. a group containing partials 2–4, and
3. a group containing partials 5–n,

where n is the highest significant partial. This *tristimulus* representation appears to be related quite closely to perceptual judgements of musical sounds.

Many instruments have noise-like qualities which strongly influence their subjective quality. A flute, for example, has a relatively simple harmonic structure, but synthetic tones with the same harmonic structure do not sound flute-like unless each note is preceded by a small puff of noise. In general, tones of standard musical instruments are poorly simulated by the summation of steady component frequencies, since such a synthesis cannot produce the dynamic variation with time characteristic of these instruments. Thus traditional electronic organs (pre-1965), which produced only tones with a fixed envelope shape, could produce a good simulation of the bagpipes, but could not be made to sound like a piano. Modern synthesizers shape the envelopes of the sounds they produce, and hence are capable of more accurate and convincing imitations of musical instruments. For a simulation to be convincing, it is often necessary to give different time envelopes to different harmonics within a complex sound [13.97, 147].

13.7 The Localization of Sounds

13.7.1 Binaural Cues

It has long been recognized that slight differences in the sounds reaching the two ears can be used as cues in sound localization. The two major cues are differences in the time of arrival at the two ears and differences in intensity at the two ears. For example, a sound coming from the left arrives first at the left ear and is more intense in the left ear. For steady sinusoids, a difference in time of arrival is equivalent to a phase difference between the sounds at the two ears. However, phase differences are not usable over the whole audible frequency range. Experiments using sounds delivered by headphones have shown that a phase difference at the two ears can be detected and used to judge location only for frequencies below about 1500 Hz. This is reasonable because, at high frequencies, the wavelength of sound is small compared to the dimensions of the head, so the listener cannot determine which cycle in the left ear corresponds to a given cycle in the right; there may be many cycles of phase difference. Thus phase differences become ambiguous and unusable at high frequencies. On the other hand, at low frequencies our accuracy at detecting changes in relative time at the two ears is remarkably good; changes of 10–20 μs can be detected, which is equivalent to a movement of the sound source of 1–2° laterally when the sound comes from straight ahead [13.148].

Intensity differences are usually largest at high frequencies. This is because low frequencies bend or diffract around the head, so that there is little difference in intensity at the two ears whatever the location of the sound source, except when the source is very close to the head. At high frequencies, the head casts more of a shadow, and above 2–3 kHz the intensity differences provide useful cues. For complex sounds, containing a range of frequencies, the difference in spectral patterning at the two ears may also be important.

The idea that sound localization is based on interaural time differences at low frequencies and interaural intensity differences at high frequencies has been called the duplex theory of sound localization, proposed by Lord *Rayleigh* [13.149]. However, complex sounds, containing only high frequencies (above 1500 Hz), can be localized on the basis of interaural time delays, provided that they have an appropriate temporal structure. For example, a single click can be localized in this way no matter what its frequency content. Periodic sounds containing only high-frequency harmonics can also be localized on the basis of interaural time differences, provided that the envelope repetition rate (usually equal to the fundamental frequency) is below about 600 Hz [13.150, 151]. Since most of the sounds we encounter in everyday life are complex, and have repetition rates below 600 Hz, interaural time differences can be used for localization in most listening situations.

13.7.2 The Role of the Pinna and Torso

Binaural cues are not sufficient to account for all aspects of sound localization. For example, an interaural time or intensity difference will not indicate whether a sound is coming from in front or behind, or above or below, but such judgments can clearly be made. Also, under some conditions localization with one ear can be as accurate as with two. It has been shown that reflections of sounds from the pinnae and torso play an important role in sound localization [13.152, 153]. The spectra of sounds entering the ear are modified by these reflections in a way which depends upon the direction of the sound source. This direction-dependent filtering provides cues for sound source location. The spectral cues are important not just in providing information about the direction of sound sources, but also in enabling us to judge whether a sound comes from within the head or from the outside world. The pinnae alter the sound spectrum primarily at high frequencies. Only when the wavelength of the sound is comparable with or smaller than the dimensions of the pinnae is the spectrum significantly affected. This occurs mostly above about 6 kHz. Thus, cues provided by the pinnae are most effective for broadband high-frequency sounds. However, reflections from other structures, such as the shoulders, result in spectral changes at lower frequencies, and these may be important for front–back discrimination [13.154].

13.7.3 The Precedence Effect

In everyday conditions the sound from a given source reaches the ears by many different paths. Some of it arrives via a direct path, but a great deal may only reach the ears after reflections from one or more surfaces. However, listeners are not normally aware of these reflections, and the reflections do not markedly impair the ability to localize sound sources. The reason for this seems to lie in a phenomenon known as the precedence effect [13.155, 156]; for a review, see [13.157]. When several sounds reach the ears in close succession (i. e., the direct sound and its reflections) the sounds are perceptually fused into a single sound (an effect called echo suppression), and the location of the total sound is primarily determined by the location of the first (direct) sound (the precedence effect). Thus the reflections have little influence on the perception of direction. Furthermore, there is little awareness of the reflections, although they may influence the timbre and loudness of the sound.

The precedence effect only occurs for sounds of a discontinuous or transient character, such as speech or music, and it can break down if the reflections have a level 10 dB or more above that of the direct sound. However, in normal conditions the precedence effect plays an important role in the localization and identification of sounds in reverberant conditions.

13.8 Auditory Scene Analysis

It is hardly ever the case that the sound reaching our ears comes from a single source. Usually the sound arises from several different sources. However, usually we are able to decompose the mixture of sounds and to perceive each source separately. An auditory object can be defined as the percept of a group of successive and/or simultaneous sound elements as a coherent whole, appearing to emanate from a single source.

As discussed earlier, the peripheral auditory system acts as a frequency analyzer, separating the different frequency components in a complex sound. Somewhere in the brain, the internal representations of these frequency components have to be assigned to their appropriate sources. If the input comes from two sources, A and B, then the frequency components must be split into two groups; the components emanating from source A should be assigned to one source and the components emanating from source B should be assigned to another. The process of doing this is often called perceptual grouping. It is also given the name auditory scene analysis [13.158]. The process of separating the elements arising from two or more different sources is sometimes called segregation.

Many different physical cues may be used to derive separate perceptual objects corresponding to the individual sources which give rise to a complex acoustic input. There are two aspects to this process: the grouping together of all the simultaneous frequency components that emanate from a single source at a given moment, and the connecting over time of the changing frequencies that a single source produces from one moment to the next [13.159]. These two aspects are sometimes described as *simultaneous grouping* and *sequential grouping*, respectively.

Most experiments on perceptual grouping have studied the effect of grouping on one specific attribute of sounds, for example, their pitch, their subjective location, or their timbre. These experiments have shown that a cue which is effective for one attribute may be less effective or completely ineffective for another attribute [13.160]. Also, the effectiveness of the cues may differ for simultaneous and sequential grouping.

13.8.1 Information Used to Separate Auditory Objects

Fundamental Frequency and Spectral Regularity

When we listen to two steady complex tones together (e.g., two musical instruments or two vowel sounds), we do not generally confuse which harmonics belong to which tone. If the complex tones overlap spectrally, two sounds are heard only if the two tones have different values of fundamental frequency (F_0) [13.161]. *Scheffers* [13.162] has shown that, if two vowels are presented simultaneously, they can be identified better when they have F_0s that differ by more than 6% than when they have the same F_0. Other researchers have reported similar findings [13.163, 164].

F_0 may be important in several ways. The components in a periodic sound have frequencies which form a simple harmonic series; the frequencies are integer multiples of F_0. This property is referred to as harmonicity. The lower harmonics are resolved in the peripheral auditory system. The regular spacing of the lower harmonics may promote their perceptual fusion, causing them to be heard as a single sound. If a sinusoidal component does not form part of this harmonic series, it tends to be heard as a separate sound. This is illustrated by some experiments of *Moore* et al.[13.165]. They investigated the effect of mistuning a single low harmonic in a harmonic complex tone. When the harmonic was mistuned sufficiently, it was heard as a separate pure tone standing out from the complex as a whole. The degree of mistuning required varied somewhat with the duration of the sounds; for 400 ms tones, a mistuning of 3% was sufficient to make the harmonic stand out as a separate tone.

Roberts and *Brunstrom* [13.166, 167] have suggested that the important feature determining whether a group of frequency components sounds fused is not harmonicity, but spectral regularity; if a group of components form a regular spectral pattern, they tend to be heard as fused, while if a single component does not fit the pattern, it is heard to pop out. For example, a sequence of components with frequencies 623, 823, 1023, 1223, and 1423 Hz is heard as relatively fused. If the frequency of the middle component is shifted to, say, 923 or 1123 Hz, that component no longer forms part of the regular pattern, and it tends to be heard as a separate tone, standing out from the complex.

For the higher harmonics in a complex sound, F_0 may play a different role. The higher harmonics of a periodic complex sound are not resolved on the basilar membrane, but give rise to a complex waveform with a periodicity equal to F_0 (see Fig. 13.18). When two complex sounds with different F_0s are presented simultaneously, then each will give rise to a waveform on the basilar membrane with periodicity equal to its respective F_0. If the two sounds have different spectra, then each will dominate the response at certain points on the basilar membrane. The auditory system may group together regions with common F_0 and segregate them from regions with different F_0 [13.163]. It may also be the case that both resolved and unresolved components can be grouped on the basis of the detailed time pattern of the neural spikes [13.168].

This process can be explained in a qualitative way by extending the model of pitch perception presented earlier. Assume that the pitch of a complex tone results from a correlation or comparison of time intervals between successive nerve firings in neurons with different CFs. Only those channels which show a high correlation would be classified as belonging to the same sound. Such a mechanism would automatically group together components with a common F_0. However, *de Cheveigné* et al. [13.169] presented evidence against such a mechanism. They measured the identification of a target vowel in the presence of a background vowel; the nominal fundamental frequencies of the two vowels differed by 6.45%. Identification was better when the background was harmonic than when it was made inharmonic (by shifting the frequency of each harmonic by a random amount between 0 and 6.45% or by less than half of the spacing between harmonics, whichever was smaller). In contrast, identification of the target did not depend upon whether or not the target was harmonic. *De Cheveigné* and coworkers [13.169–171] proposed a mechanism based on the idea that a harmonic background sound can be canceled in the auditory system, thus enhancing the representation of a target vowel.

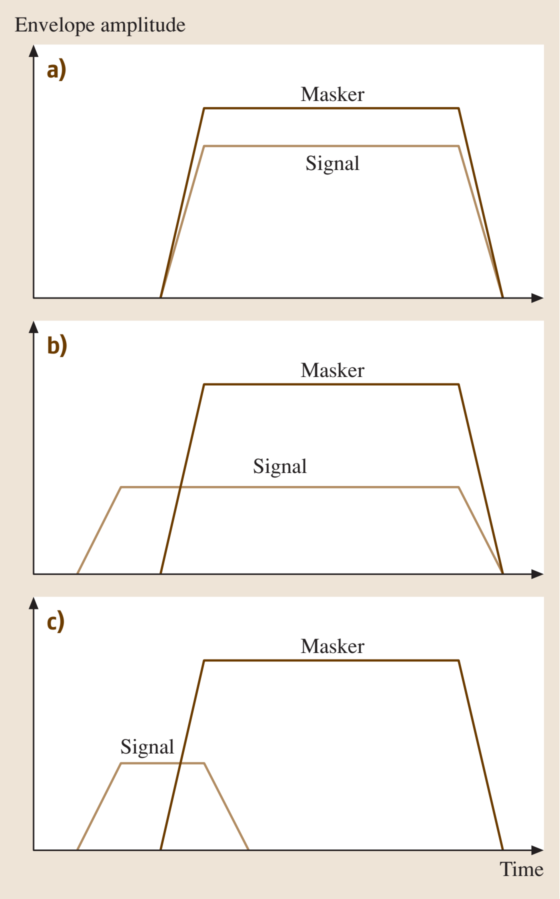

Fig. 13.19a–c Schematic illustration of the stimuli used by *Rasch* [13.172]. Both the signal and the masker were periodic complex tones, with the signal having the higher fundamental frequency. When the signal and masker were gated on and off synchronously **(a)**, the threshold for the signal was relatively high. When the signal started slightly before the masker **(b)**, the threshold was markedly reduced. When the signal was turned off as soon as the masker was turned on **(c)**, the signal was perceived as continuing through the masker, and the threshold was the same as when the signal did continue through the masker

Onset and Offset Disparities

Another cue for the perceptual separation of (near-)simultaneous sounds is onset and offset disparity. *Rasch* [13.172] investigated the ability to hear one complex tone in the presence of another. One of the tones was treated as a masker and the level of the signal tone (the higher in F_0) was adjusted to find the point where it was just detectable. When the two tones started at the same time and had exactly the same temporal envelope, the threshold of the signal was between 0 and −20 dB relative to the level of the masker (Fig. 13.19a). Thus, when a difference in F_0 was the only cue, the signal could not be heard when its level was more than 20 dB below that of the masker.

Rasch also investigated the effect of starting the signal just before the masker (Fig. 13.19b). He found that threshold depended strongly on onset asynchrony, reaching a value of −60 dB for an asynchrony of 30 ms. Thus, when the signal started 30 ms before the masker, it could be heard much more easily and with much greater differences in level between the two tones. It should be emphasized that the lower threshold was a result of the signal occurring for a brief time on its own; essentially performance was limited by backward masking of the 30 ms asynchronous segment, rather than by simultaneous masking. However, the experiment does illustrate the large benefit that can be obtained from a relatively small asynchrony.

Although the percept of his subjects was that the signal continued throughout the masker, Rasch showed that this percept was not based upon sensory information received during the presentation time of the masker. He found that identical thresholds were obtained if the signal terminated immediately after the onset of the masker (Fig. 13.19c). It appears that the perceptual system "assumes" that the signal continues, since there is no evidence to the contrary; the part of the signal that occurs simultaneously with the masker would be completely masked.

Rasch [13.172] showed that, if the two tones have simultaneous onsets but different rise times, this also can give very low thresholds for the signal, provided it has the shorter rise time. Under these conditions and those of onset asynchronies up to 30 ms, the notes sound as though they start synchronously. Thus, we do not need to be consciously aware of the onset differences for the auditory system to be able to exploit them in the perceptual separation of complex tones. Rasch also pointed out that, in ensemble music, different musicians do not play exactly in synchrony even if the score indicates that they should. The onset differences used in his experiments correspond roughly with the onset asynchronies of nominally simultaneous notes found in performed music. This supports the view that the asynchronies are an important factor in the perception of the separate parts or voices in polyphonic music.

Onset asynchronies can also play a role in determining the timbre of complex sounds. *Darwin* and *Sutherland* [13.173] showed that a tone that starts or stops at a different time from a vowel is less likely to be heard as part of that vowel than if it is simultaneous with it. For example, increasing the level of a single harmonic can produce a significant change in the quality (timbre) of a vowel. However, if the incremented harmonic starts before the vowel, the change in vowel quality is markedly reduced. Similarly, *Roberts* and *Moore* [13.174] showed that extraneous sinusoidal components added to a vowel could influence vowel quality, but the influence was markedly reduced when the extraneous components were turned on before the vowel or turned off after the vowel.

Contrast with Previous Sounds

The auditory system seems well suited to the analysis of changes in the sensory input, and particularly to changes in spectrum over time. The changed aspect stands out perceptually from the rest. It is possible that there are specialized central mechanisms for detecting changes in spectrum. Additionally, stimulation with a steady sound may result in some kind of adaptation. When some aspect of a stimulus is changed, that aspect is freed from the effects of adaptation and thus will be enhanced perceptually. While the underlying mechanism is a matter of debate, the perceptual effect certainly is not.

A powerful demonstration of this effect may be obtained by listening to a stimulus with a particular spectral structure and then switching rapidly to a stimulus with a flat spectrum, such as white noise. A white noise heard in isolation may be described as colorless; it has no pitch and has a neutral sort of timbre. However, when a white noise follows soon after a stimulus with spectral structure, the noise sounds colored [13.175]. A harmonic complex tone with a flat spectrum may be given a speech-like quality if it is preceded by a harmonic complex having a spectrum which is the inverse of that of a speech sound, such as a vowel [13.176].

Another demonstration of the effects of a change in a stimulus can be obtained by listening to a steady complex tone with many harmonics. Usually such a tone is heard with a single pitch corresponding to F_0, and the individual harmonics are not separately perceived. However, if one of the harmonics is changed in some way, by altering either its relative phase or its level, then that harmonic stands out perceptually from the complex as a whole. For a short time after the change is made, a pure-tone quality is perceived. The perception of the harmonic then gradually fades, until it merges with the complex once more.

Change detection is obviously of importance in assigning sound components to their appropriate sources. Normally we listen against a background of sounds which may be relatively unchanging, such as the humming of machinery, traffic noises, and so on. A sudden change in the sound is usually indicative that a new source has been activated, and the change detection mechanisms enable us to isolate the effects of the change and interpret them appropriately.

Correlated Changes in Amplitude or Frequency

Rasch [13.172] also showed that, when the two complex tones start and end synchronously, the detection of the tone with the higher F_0 could be enhanced by frequency modulating it. The modulation was similar to the vibrato which often occurs for musical tones, and it was applied so that all the components in the higher tone moved up and down in synchrony. Rasch found that the modulation could reduce the threshold for detecting the higher tone by 17 dB. A similar effect can be produced by amplitude modulation (AM) of one of the tones. The modulation seems to enhance the salience of the modulated sound, making it appear to stand out from the unmodulated sound.

It is less clear whether the perceptual segregation of simultaneous sounds is affected by the coherence of changes in amplitude or frequency when *both* sounds are modulated. Coherence here refers to whether the changes in amplitude or frequency of the two sounds have the same pattern over time or different patterns over time. Several experiments have been reported suggesting that coherence of amplitude changes plays a role; sounds with coherent changes tend to fuse perceptually, whereas sounds with incoherent changes tend to segregate [13.177–180]. However, *Summerfield* and *Culling* [13.181] found that the coherence of AM did not affect the identification of pairs of simultaneous vowels when the vowels were composed of components placed randomly in frequency (to avoid effects of harmonicity).

Evidence for a role of frequency modulation (FM) coherence in perceptual grouping has been more elusive. While some studies have been interpreted as indicating a weak role for frequency modulation coherence [13.182, 183], the majority of studies have failed to indicate such sensitivity [13.181, 184–188]. *Furukawa* and *Moore* [13.189] have shown that the detectability of FM imposed on two widely separated carrier frequencies is better when the modulation is coherent on the two carriers than when it is incoherent. However, this may occur because the overall pitch evoked by the two carriers fluctuates more when the carriers are modulated coherently than when they are modulated incoherently [13.190–192]. There is at present no clear evidence that the coherence of FM influences perceptual grouping when both sounds are modulated.

Sound Location

The cues used in sound localization may also help in the analysis of complex auditory inputs. A phenomenon that is related to this is called the binaural masking level difference (MLD). The phenomenon can be summarized as follows: whenever the phase or level differences of a signal at the two ears are not the same as those of a masker, the ability to detect the signal is improved relative to the case where the signal and masker have the same phase and level relationships at the two ears. The practical implication is that a signal is easier to detect when it is located in a different position in space from the masker. Although most studies of the MLD have been concerned with threshold measurements, it seems clear that similar advantages of binaural listening can be gained in the identification and discrimination of signals presented against a background of other sound.

An example of the use of binaural cues in separating an auditory object from its background comes from an experiment by *Kubovy* et al. [13.193]. They presented eight continuous sinusoids to each ear via earphones. The sinusoids had frequencies corresponding to the notes in a musical scale, the lowest having a frequency of 300 Hz. The input to one ear, say the left, was presented with a delay of 1 ms relative to the input to the other ear, so the sinusoids were all heard toward the right side of the head. Then, the phase of one of the sinusoids was advanced in the left ear, while its phase in the right ear was delayed, until the input to the left ear led the input to the right ear by 1 ms; this phase-shifting process occurred over a time of 45 ms. The phase remained at the shifted value for a certain time and was then smoothly returned to its original value, again over 45 ms. During the time that the phase shift was present, the phase-shifted sinusoid appeared toward the opposite (left) side of the head, making it stand out perceptually. A sequence of phase shifts in different components was clearly heard as a melody. This melody was completely undetectable when listening to the input to one ear alone. Kubovy et al. interpreted their results as indicating that differences in relative phase at the two ears can allow an auditory object to be isolated in the absence of any other cues.

Culling [13.194] performed a similar experiment to that of *Kubovy* et al. [13.193], but he examined the importance of the phase transitions. He found that, when one component of a complex sound changed rapidly but smoothly in interaural time difference (ITD), it perceptually segregated from the complex. When different components were changed in ITD in succession, a recognizable melody was heard, as reported by Kubovy et al. However, when the transitions were replaced by silent intervals, leaving only static ITDs as a cue, the melody was much less salient. Thus, transitions in ITD seem to be more important than static differences in ITD in producing segregation of one component from a background of other components. Nevertheless, static differences in ITD do seem to be sufficient to produce segregation under some conditions [13.195].

Other experiments suggest that binaural processing often plays a relatively minor role in simultaneous grouping. *Shackleton* and *Meddis* [13.196] investigated the ability to identify each vowel in pairs of concurrent vowels. They found that a difference in F_0 between the two vowels improved identification by about 22%. In contrast, a 400 μs interaural delay in one vowel (which corresponds to an azimuth of about 45 degrees) improved performance by only 7%. *Culling* and *Summerfield* [13.197] investigated the identification of concurrent whispered vowels, synthesized using bands of noise. They showed that listeners were able to identify the vowels accurately when each vowel was presented to a different ear. However, they were unable to identify the vowels when they were presented to both ears but with different ITDs. In other words, listeners could not group the noise bands in different frequency regions with the same ITD and thereby separate them from noise bands in other frequency regions with a different ITD.

In summary, when two simultaneous sounds differ in their interaural level or time, this can contribute to the perceptual segregation of the sounds and enhance their detection and discrimination. However, such binaural processing is not always effective, and in some situations it appears to play little role.

13.8.2 The Perception of Sequences of Sounds

Stream Segregation

When we listen to rapid sequences of sounds, the sounds may be grouped together (i. e., perceived as if they come from a single source, called fusion or coherence), or they may be perceived as different streams (i. e., as coming from more than one source, called fission or stream segregation) [13.158, 199–201]. The term streaming is used to denote the processes determining whether one stream or multiple streams are heard. *Van Noorden* [13.202] investigated this phenomenon using a sequence of pure tones where every second B was omitted from the regular sequence ABABAB . . . , producing a sequence ABA ABA . . . . He found that this could be perceived in two ways, depending on the frequency separation of A and B. For small separations, a single rhythm, resembling a gallop, is heard (fusion). For larger separations, two separate tone sequences can be heard, one of which (A A A) is running twice as fast as the other (B B B) (fission). Components are more likely to be assigned to separate streams if they differ widely in frequency or if there are rapid jumps in frequency between them. The latter point is illustrated by a study of *Bregman* and *Dannenbring* [13.203]. They used tone sequences in which successive tones were connected by frequency glides. They found that these glides reduced the tendency for the sequences to split into high and low streams. Conditions using partial glides also showed decreased stream segregation, although the partial glides were not quite as effective as complete glides. Thus, complete continuity between tones is not required to reduce stream segregation; a frequency change pointing toward the next tone allows the listener to follow the pattern more easily.

The effects of frequency glides and other types of transitions in preventing stream segregation or fission are probably of considerable importance in the perception of speech. Speech sounds may follow one another in very rapid sequences, and the glides and partial glides observed in the acoustic components of speech may be a strong factor in maintaining the percept of speech as a unified stream.

Van Noorden found that, for intermediate frequency separations of the tones A and B in a rapid sequence, either fusion or fission could be heard, according to the instructions given and the attentional set of the subject. When the percept is ambiguous, the tendency for fission to occur increases with increasing exposure time to the tone sequence [13.204]. The auditory system seems to start with the assumption that there is a single sound source, and fission is only perceived when sufficient evidence has built up to contradict this assumption. Sudden changes in a sequence, or in the perception of a sequence, can cause the percept to revert to its initial, default condition, which is fusion [13.205, 206]; for a review, see [13.207].

For rapid sequences of complex tones, strong fission can be produced by differences in spectrum of successive tones, even when all tones have the same F_0 [13.201, 208–210]. However, when successive complex tones are filtered to have the same spectral envelope, stream segregation can also be produced by differences between successive tones in F_0 [13.210, 211], in temporal envelope [13.209, 212], in the relative phases of the components [13.213], or in apparent location [13.214]. *Moore* and *Gockel* [13.207] proposed that any salient perceptual difference between successive tones may lead to stream segregation. Consistent with this idea, *Dowling* [13.215, 216] has shown that stream segregation may also occur when successive pure tones differ in intensity or in spatial location. He presented a melody composed of equal-intensity notes and inserted between each note of the melody a tone of the same intensity, with a frequency randomly selected from the same range. He found that the resulting tone sequence produced a meaningless jumble. Making the interposed notes different from those of the melody, in either intensity, frequency range, or spatial location, caused them to be heard as a separate stream, enabling subjects to pick out the melody.

Darwin and *Hukin* [13.198] have shown that sequential grouping can be strongly influenced by ITD. In one experiment, they simultaneously presented two

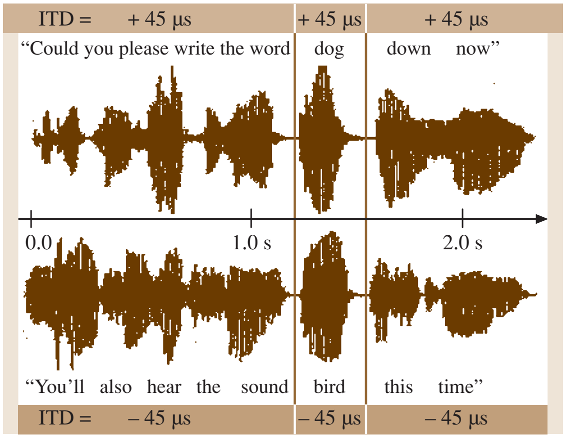

Fig. 13.20 Example of the stimuli used by *Darwin* and *Hukin* [13.198] (After Fig. 1 of [13.198])

sentences (see Fig. 13.20). They varied the ITDs of the two sentences in the range 0 to ±181 μ s. For example, one sentence might lead in the left ear by 45 μ s, while the other sentence would lead in the right ear by 45 μ s (as in Fig. 13.20). The sentences were based on natural speech but were processed so that each was spoken on a monotone, i. e., with constant F_0. The F_0 difference between the two sentences was varied from 0 to 4 semitones. Subjects were instructed to attend to one particular sentence. At a certain point, the two sentences contained two different target words aligned in starting time and duration ("dog" and "bird"). The F_0s and the ITDs of the two target words were varied independently from those of the two sentences. Subjects had to indicate which of the two target words they heard in the attended sentence. They reported the target word that had the same ITD as the attended sentence much more often than the target word with the opposite ITD. In other words, the target word with the same ITD as the attended sentence was grouped with that sentence. This was true even when the target word had the same ITD as the attended sentence but a different F_0. Thus, subjects grouped words across time according to their perceived location, independent of F_0 differences. *Darwin* and *Hukin* [13.198] concluded that listeners who try to track a particular sound source over time direct attention to auditory objects at a particular subjective location. The auditory objects themselves may be formed using cues other than ITD, for example, onset and offset asynchrony and harmonicity.

It should be noted that, for discrete sequences of musical tones, the auditory system does not necessarily form streams according to perceived location, especially when that cue competes with other cues. This is illustrated by an effect, called the scale illusion, reported by *Deutsch* [13.217]. She presented two sequences of tones via headphones, one sequence to each ear. The *n*th tone in the left ear was synchronous with the *n*th tone in the right ear. The sequences were created by repetitive presentation of the C major scale in both ascending and descending form, such that when a component of the ascending scale was in one ear, a component of the descending scale was in the other, and vice versa. However, the tones from each scale alternated between ears. Within each ear there were often large jumps in frequency between successive tones. Most subjects perceived the sounds as two streams, organized by the frequency proximity of successive tones. One stream (which was often heard towards one ear) was heard as a musical scale that started high, descended and then increased again, while the other stream (which was usually heard towards the opposite ear) was heard as a scale that started low, ascended, and then decreased again. Thus, the true location of the tones had little influence on the formation of the perceptual streams.

Another example come from the opening bars of the last movement of Tchaikovsky's sixth symphony. This contains interleaved notes played by the first and second violins, who according to 19th century custom sat on opposite sides of the stage. These notes are perceived as a single stream, despite the difference in location, presumably because of the frequency proximity between successive notes.

A number of composers have exploited the fact that stream segregation occurs for tones that are widely separated in frequency. By playing a sequence of tones in which alternate notes are chosen from separate frequency ranges, an instrument such as the flute, which is only capable of playing one note at a time, can appear to be playing two themes at once. Many fine examples of this are available in the works of Bach, Telemann and Vivaldi.

Judgment of Temporal Order

It is difficult to judge the temporal order of sounds that are perceived in different streams. An example of this comes from the work of *Broadbent* and *Ladefoged* [13.218]. They reported that extraneous sounds in sentences were grossly mislocated. For example, a click might be reported as occurring a word or two away from its actual position. Surprisingly poor performance was also reported by *Warren* et al. [13.219] for judgments of the temporal order of three or four unrelated items, such as a hiss, a tone, and a buzz. Most subjects could not identify the order when each successive item lasted as long as 200 ms. Naive subjects required that each item last at least 700 ms to identify the order of four sounds presented in an uninterrupted repeated sequence. These durations are well above those which are normally considered necessary for temporal resolution.

The poor order discrimination described by Warren et al. is probably a result of stream segregation. The sounds they used do not represent a coherent class. They have different temporal and spectral characteristics, and, as for tones widely differing in frequency, they do not form a single perceptual stream. Items in different streams appear to float about with respect to each other in subjective time. Thus, temporal order judgments are difficult. It should be emphasized that the relatively poor performance reported by *Warren* et al. [13.219] is found only in tasks requiring absolute identification of the order of sounds and not in tasks which simply require the discrimination of different sequences. Also, with extended training and feedback subjects can learn to distinguish between and identify orders within sequences of unrelated sounds lasting only 10 ms or less [13.220].

To explain these effects, *Divenyi* and *Hirsh* [13.221] suggested that two kinds of perceptual judgments are involved. At longer item durations the listener is able to hear a clear sequence of steady state sounds, whereas at shorter durations a change in the order of items introduces qualitative changes that can be discriminated by trained listeners. Similar explanations have been put forward by *Green* [13.75] and *Warren* [13.220].

Bregman and *Campbell* [13.200] investigated the factors that make temporal order judgments for tone sequences difficult. They used naive subjects, so performance presumably depended on the subjects actually perceiving the sounds as a sequence, rather than on their learning the overall sound pattern. They found that, in a repeating cycle of mixed high and low tones, subjects could discriminate the order of the high tones relative to one another or of the low tones among themselves, but they could not order the high tones relative to the low ones. The authors suggested that this was because the two groups of sounds split into separate perceptual streams and that judgments across streams are difficult. Several more recent studies have used tasks involving the discrimination of changes in timing or rhythm as a tool for studying stream segregation [13.210, 213, 222]. The rationale of these studies is that, if the ability to judge the relative timing of successive sound elements is good, this indicates that the elements are perceived as part of a single stream, while if the ability is poor, this indicates that the elements are perceived in different streams.

13.8.3 General Principles of Perceptual Organization

The Gestalt psychologists [13.223] described many of the factors which govern perceptual organization, and their descriptions and principles apply reasonably well to the way physical cues are used to achieve perceptual grouping of the acoustic input. It seems likely that the rules of perceptual organization have arisen because, on the whole, they tend to give the right answers. That is, use of the rules generally results in a grouping of those parts of the acoustic input that arose from the same source and a segregation of those that did not. No single rule will always work, but it appears that the rules can generally be used together, in a coordinated and probably quite complex way, in order to arrive at a correct interpretation of the input. In the following sections, I outline the major principles or rules of perceptual organization. Many, but not all, of the rules apply to both vision and hearing, and they were mostly described first in relation to vision.

Similarity

This principle is that elements will be grouped if they are similar. In hearing, similarity usually implies closeness of timbre, pitch, loudness, or subjective location. Examples of this principle have already been described. If we listen to a rapid sequence of pure tones, say 10 tones per second, then tones which are closely spaced in frequency, and are therefore similar, form a single perceptual stream, whereas tones which are widely spaced form separate streams.

For pure tones, frequency is the most important factor governing similarity, although differences in level and subjective location between successive tones can also lead to stream segregation. For complex tones, differences in timbre produced by spectral differences seem to be the most important factor. Again, however, other factors may play a role. These include differences in F_0, differences in timbre produced by temporal envelope differences, and differences in perceived location.

Good Continuation

This principle exploits a physical property of sound sources, that changes in frequency, intensity, location, or spectrum tend to be smooth and continuous, rather than abrupt. Hence, a smooth change in any of these aspects indicates a change within a single source, whereas a sudden change indicates that a new source has been activated. One example has already been described; *Bregman* and *Dannenbring* [13.203] showed that the tendency of a sequence of high and low tones to split into two streams was reduced when successive tones were connected by frequency glides.

A second example comes from studies using synthetic speech. In such speech, large fluctuations of an unexpected kind in F_0 (and correspondingly in the pitch) give the impression that a new speaker has stepped in to take over a few syllables from the primary speaker. *Darwin* and *Bethell-Fox* [13.224] synthesized spectral patterns which changed smoothly and repeatedly between two vowel sounds. When the F_0 of the sound patterns was constant they were heard as coming from a single source, and the speech sounds heard included glides ("l" as in let) and semivowels ("w" as in we). When a discontinuous, step-like F_0 contour was imposed on the patterns, they were perceived as two distinct speech streams, and the speech was perceived as containing predominantly stop consonants (e.g., "b" as in be, and "d" as in day). Apparently, a given group of components is usually only perceived as part of one stream. Thus, the perceptual segregation produces illusory silences in each stream during the portions of the signal attributed to the other stream, and these silences are interpreted, together with the gliding spectral patterns in the vowels, as indicating the presence of stop consonants. It is clear that the perception of speech sounds can be strongly influenced by stream organization.

Common Fate

The different frequency components arising from a single sound source usually vary in a highly coherent way. They tend to start and finish together, change in intensity together, and change in frequency together. This fact is exploited by the perceptual system and gives rise to the principle of common fate: if two or more components in a complex sound undergo the same kinds of changes at the same time, then they are grouped and perceived as part of the same source.

Two examples of common fate were described earlier. The first concerns the role of the onsets and offsets of sounds. Components will be grouped together if they start and stop synchronously; otherwise they will form separate streams. The onset asynchronies necessary to allow the separation of two complex tones are not large, about 30 ms being sufficient. The asynchronies which are observed in performed music are typically as large as or larger than this, so when we listen to polyphonic music we are easily able to hear separately the melodic line of each instrument. Secondly, components which are amplitude modulated in a synchronous way tend to be grouped together. There is at present little evidence that the coherence of modulation in frequency affects perceptual grouping, although frequency modulation of a group of components in a complex sound can promote the perceptual segregation of those components from an unchanging background.

Disjoint Allocation

Broadly speaking, this principle, also known as belongingness, is that a single component in a sound can only be assigned to one source at a time. In other words, once a component has been used in the formation of one stream, it cannot be used in the formation of a second stream. For certain types of stimuli, the perceptual organization may be ambiguous, there being more than one way to interpret the sensory input. When a given component might belong to one of a number of streams, the percept may alter depending on the stream within which that component is included.

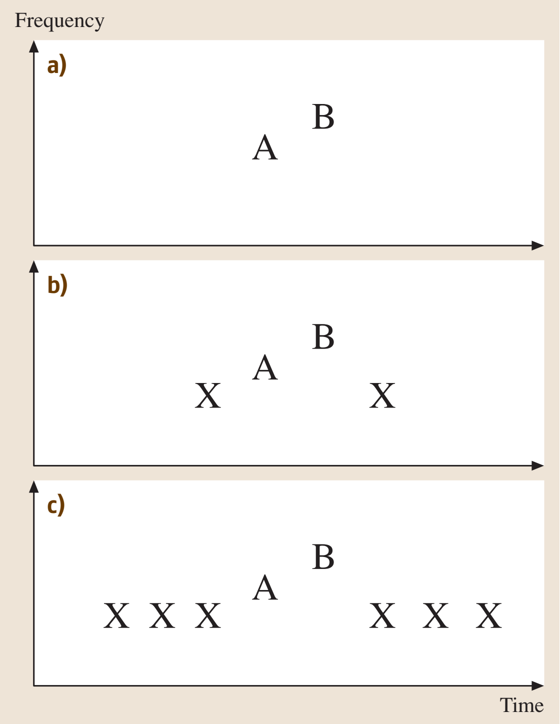

Fig. 13.21 Schematic illustration of the stimuli used by *Bregman* and *Rudnicky* [13.225]. When the tones A and B are presented alone *(panel a)*, it is easy to tell their order. When the tones A and B are presented as part of a four-tone complex XABX *(panel b)*, it is more difficult to tell their order. If the four-tone complex is embedded in a longer sequence of X tones *(panel c)*, the Xs form a separate perceptual stream, and it is easy to tell the order of A and B

An example is provided by the work of *Bregman* and *Rudnicky* [13.225]. They presented a sequence of four brief tones in rapid succession. Two of the tones, labeled X, had the same frequency, but the middle two, A and B, were different. The four-tone sequence was either XABX or XBAX. The listeners had to judge the order of A and B. This was harder than when the tones AB occurred in isolation (Fig. 13.21a) because A and B were perceived as part of a longer four-tone pattern, including the two "distracter" tones, labeled X (Fig. 13.21b). They then embedded the four-tone sequence into a longer sequence of tones, called captor tones (Fig. 13.21c). When the captor tones had frequencies close to those of the distracter tones, they captured the distracters into a separate perceptual stream, leaving the tones AB in a stream of their own. This made the order of A and B easy to judge. It seems that the tones X could not be perceived as part of both streams. When only one stream is the subject of judgment and hence of attention, the other one may serve to remove distracters from the domain of attention.

It should be noted that the principle of disjoint allocation does not always work, particularly in situations where there are two or more plausible perceptual organizations [13.226]. In such situations, a sound element may sometimes be heard as part of more than one stream.

Closure

In everyday life, the sound from a given source may be temporarily masked by other sounds. While the masking sound is present there may be no sensory evidence which can be used to determine whether the masked sound has continued or not. Under these conditions the masked sound tends to be perceived as continuous. The Gestalt psychologists called this process closure.

A laboratory example of this phenomenon is the continuity effect [13.227–229]. When a sound A is alternated with a sound B, and B is more intense than A, then A may be heard as continuous, even though it is interrupted. The sounds do not have to be steady. For example, if B is noise and A is a tone which is gliding upward in frequency, the glide is heard as continuous even though certain parts of the glide are missing [13.230]. Notice that, for this to be the case, the gaps in the tone must be filled with noise and the noise must be a potential masker of the tone (if they were presented simultaneously). In the absence of noise, discrete jumps in frequency are clearly heard.

The continuity effect also works for speech stimuli alternated with noise. In the absence of noise to fill in the gaps, interrupted speech sounds hoarse and raucous. When noise is presented in the gaps, the speech sounds more natural and continuous [13.231]. For connected speech at moderate interruption rates, the intervening noise actually leads to an improvement in intelligibility [13.232]. This may occur because the abrupt switching of the speech produces misleading cues as to which speech sounds were present. The noise serves to mask these misleading cues.

It is clear from these examples that the perceptual filling in of missing sounds does not take place solely on the basis of evidence in the acoustic waveform. Our past experience with speech, music, and other stimuli must play a role, and the context of surrounding sounds is important [13.233]. However, the filling in only occurs when one source is perceived as masking or occluding another. This percept must be based on acoustic evidence that the occluded sound has been masked. Thus, if a gap is not filled by a noise or other sound, the perceptual closure does not occur; a gap is heard.

13.9 Further Reading and Supplementary Materials

More information about the topics discussed in this chapter can be found in: A. S. Bregman: *Auditory Scene Analysis: The Perceptual Organization of Sound* (Bradford Books, MIT Press, Cambridge 1990); W. M. Hartmann: *Signals, Sound, and Sensation* (AIP Press, Woodbury 1997); B. C. J. Moore: *An Introduction to the Psychology of Hearing, 5th Ed.* (Academic, San Diego 2003); R. Plomp: *The Intelligent Ear* (Erlbaum, Mahwah 2002)

A compact disc (CD) of auditory demonstrations has been produced by A. J. M. Houtsma, T. D. Rossing, W. M. Wagenaars (1987). The disc can be obtained through the Acoustical Society of America; contact asapubs@abdintl.com for details.

The following CD has a large variety of demonstrations relevant to perceptual grouping: A. S. Bregman, P. Ahad (1995). *Demonstrations of Auditory Scene Analysis: The Perceptual Organization of Sound*, (Auditory Perception Laboratory, Department of Psychology, McGill University, distributed by MIT Press, Cambridge, MA). It can be ordered from The MIT Press, 55 Hayward St., Cambridge, MA 02142, USA. Further relevant demonstrations can be found at: http://www.kyushu-id.ac.jp/∼ynhome/.

A CD simulating the effects of a hearing loss on the perception of speech and music is *Perceptual Consequences of Cochlear Damage*, which may be obtained by writing to B. C. J. Moore, Department of Experimental Psychology, University of Cambridge, Downing Street, Cambridge, CB2 3EB, England, and enclosing a check for 20 dollars or a cheque for 12 pounds sterling, made payable to B. C. J. Moore.

References

13.1 ISO 389-7: *Acoustics – Reference zero for the calibration of audiometric equipment. Part 7: Reference threshold of hearing under free-field and diffuse-field listening conditions* (International Organization for Standardization, Geneva 1996)

13.2 M.C. Killion: Revised estimate of minimal audible pressure: Where is the "missing 6 dB"?, J. Acoust. Soc. Am. **63**, 1501–1510 (1978)

13.3 B.C.J. Moore, B.R. Glasberg, T. Baer: A model for the prediction of thresholds, loudness and partial loudness, J. Audio Eng. Soc. **45**, 224–240 (1997)

13.4 M.A. Cheatham, P. Dallos: Inner hair cell response patterns: implications for low-frequency hearing, J. Acoust. Soc. Am. **110**, 2034–2044 (2001)

13.5 L.J. Sivian, S.D. White: On minimum audible sound fields, J. Acoust. Soc. Am. **4**, 288–321 (1933)

13.6 I. Pollack: Monaural and binaural threshold sensitivity for tones and for white noise, J. Acoust. Soc. Am. **20**, 52–57 (1948)

13.7 J.K. Dierks, L.A. Jeffress: Interaural phase and the absolute threshold for tone, J. Acoust. Soc. Am. **34**, 981–986 (1962)

13.8 K. Krumbholz, R.D. Patterson, D. Pressnitzer: The lower limit of pitch as determined by rate discrimination, J. Acoust. Soc. Am. **108**, 1170–1180 (2000)

13.9 R. Plomp, M.A. Bouman: Relation between hearing threshold and duration for tone pulses, J. Acoust. Soc. Am. **31**, 749–758 (1959)

13.10 ANSI: *ANSI S1.1-1994. American National Standard Acoustical Terminology* (American National Standards Institute, New York 1994)

13.11 R.L. Wegel, C.E. Lane: The auditory masking of one sound by another and its probable relation to the dynamics of the inner ear, Phys. Rev. **23**, 266–285 (1924)

13.12 L.L. Vogten: Low-level pure-tone masking: a comparison of 'tuning curves' obtained with simultaneous and forward masking, J. Acoust. Soc. Am. **63**, 1520–1527 (1978)

13.13 H. Fletcher: Auditory patterns, Rev. Mod. Phys. **12**, 47–65 (1940)

13.14 H.L.F. Helmholtz: *Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik* (Vieweg, Braunschweig 1863)

13.15 R.D. Patterson, B.C.J. Moore: Auditory filters and excitation patterns as representations of frequency resolution. In: *Frequency Selectivity in Hearing*, ed. by B.C.J. Moore (Academic, London 1986)

13.16 D. Johnson-Davies, R.D. Patterson: Psychophysical tuning curves: restricting the listening band to the signal region, J. Acoust. Soc. Am. **65**, 765–770 (1979)

13.17 B.J. O'Loughlin, B.C.J. Moore: Off-frequency listening: effects on psychoacoustical tuning curves obtained in simultaneous and forward masking, J. Acoust. Soc. Am. **69**, 1119–1125 (1981)

13.18 B.J. O'Loughlin, B.C.J. Moore: Improving psychoacoustical tuning curves, Hear. Res. **5**, 343–346 (1981)

13.19 R.D. Patterson: Auditory filter shapes derived with noise stimuli, J. Acoust. Soc. Am. **59**, 640–654 (1976)

13.20 B.R. Glasberg, B.C.J. Moore: Derivation of auditory filter shapes from notched-noise data, Hear. Res. **47**, 103–138 (1990)

13.21 J.P. Egan, H.W. Hake: On the masking pattern of a simple auditory stimulus, J. Acoust. Soc. Am. **22**, 622–630 (1950)

13.22 E. Zwicker, E. Terhardt: Analytical expressions for critical band rate and critical bandwidth as a function of frequency, J. Acoust. Soc. Am. **68**, 1523–1525 (1980)

13.23 R.D. Patterson, I. Nimmo-Smith: Off-frequency listening and auditory filter asymmetry, J. Acoust. Soc. Am. **67**, 229–245 (1980)

13.24 B.C.J. Moore, B.R. Glasberg: Formulae describing frequency selectivity as a function of frequency and level and their use in calculating excitation patterns, Hear. Res. **28**, 209–225 (1987)

13.25 S. Rosen, R.J. Baker, A. Darling: Auditory filter nonlinearity at 2 kHz in normal hearing listeners, J. Acoust. Soc. Am. **103**, 2539–2550 (1998)

13.26 B.R. Glasberg, B.C.J. Moore, R.D. Patterson, I. Nimmo-Smith: Dynamic range and asymmetry of the auditory filter, J. Acoust. Soc. Am. **76**, 419–427 (1984)

13.27 E. Zwicker, H. Fastl: *Psychoacoustics – Facts and Models*, 2nd edn. (Springer, Berlin 1999)

13.28 B.C.J. Moore, J.I. Alcántara, T. Dau: Masking patterns for sinusoidal and narrowband noise maskers, J. Acoust. Soc. Am. **104**, 1023–1038 (1998)

13.29 B.C.J. Moore, B.R. Glasberg: Suggested formulae for calculating auditory-filter bandwidths and excitation patterns, J. Acoust. Soc. Am. **74**, 750–753 (1983)

13.30 K. Miyazaki, T. Sasaki: Pure-tone masking patterns in nonsimultaneous masking conditions, Jap. Psychol. Res. **26**, 110–119 (1984)

13.31 A.J. Oxenham, B.C.J. Moore: Modeling the additivity of nonsimultaneous masking, Hear. Res. **80**, 105–118 (1994)

13.32 B.C.J. Moore, B.R. Glasberg: Growth of forward masking for sinusoidal and noise maskers as a function of signal delay: implications for suppression in noise, J. Acoust. Soc. Am. **73**, 1249–1259 (1983)

13.33 G. Kidd, L.L. Feth: Effects of masker duration in pure-tone forward masking, J. Acoust. Soc. Am. **72**, 1384–1386 (1982)

13.34 E. Zwicker: Dependence of post-masking on masker duration and its relation to temporal effects in loudness, J. Acoust. Soc. Am. **75**, 219–223 (1984)

13.35 H. Fastl: Temporal masking effects: I. Broad band noise masker, Acustica **35**, 287–302 (1976)

13.36 H. Duifhuis: Audibility of high harmonics in a periodic pulse II. Time effects, J. Acoust. Soc. Am. **49**, 1155–1162 (1971)

13.37 H. Duifhuis: Consequences of peripheral frequency selectivity for nonsimultaneous masking, J. Acoust. Soc. Am. **54**, 1471–1488 (1973)

13.38 C.J. Plack, B.C.J. Moore: Temporal window shape as a function of frequency and level, J. Acoust. Soc. Am. **87**, 2178–2187 (1990)

13.39 W. Jesteadt, S.P. Bacon, J.R. Lehman: Forward masking as a function of frequency, masker level, and signal delay, J. Acoust. Soc. Am. **71**, 950–962 (1982)

13.40 R.L. Smith: Short-term adaptation in single auditory-nerve fibers: Some poststimulatory effects, J. Neurophysiol. **49**, 1098–1112 (1977)

13.41 C.W. Turner, E.M. Relkin, J. Doucet: Psychophysical and physiological forward masking studies: Probe duration and rise-time effects, J. Acoust. Soc. Am. **96**, 795–800 (1994)

13.42 A.J. Oxenham: Forward masking: adaptation or integration?, J. Acoust. Soc. Am. **109**, 732–741 (2001)

13.43 M. Brosch, C.E. Schreiner: Time course of forward masking tuning curves in cat primary auditory cortex, J. Neurophysiol. **77**, 923–943 (1997)

13.44 A.J. Oxenham, B.C.J. Moore: Additivity of masking in normally hearing and hearing-impaired subjects, J. Acoust. Soc. Am. **98**, 1921–1934 (1995)

13.45 R. Plomp: The ear as a frequency analyzer, J. Acoust. Soc. Am. **36**, 1628–1636 (1964)

13.46 R. Plomp, A.M. Mimpen: The ear as a frequency analyzer II, J. Acoust. Soc. Am. **43**, 764–767 (1968)

13.47 B.C.J. Moore, K. Ohgushi: Audibility of partials in inharmonic complex tones, J. Acoust. Soc. Am. **93**, 452–461 (1993)

13.48 D.R. Soderquist: Frequency analysis and the critical band, Psychon. Sci. **21**, 117–119 (1970)

13.49 P.A. Fine, B.C.J. Moore: Frequency analysis and musical ability, Music Percept. **11**, 39–53 (1993)

13.50 B. Gabriel, B. Kollmeier, V. Mellert: Influence of individual listener, measurement room and choice of test-tone levels on the shape of equal-loudness level contours, Acustica Acta Acustica **83**, 670–683 (1997)

13.51 D. Laming: *The Measurement of Sensation* (Oxford University Press, Oxford 1997)

13.52 H. Fletcher, W.A. Munson: Loudness, its definition, measurement and calculation, J. Acoust. Soc. Am. **5**, 82–108 (1933)

13.53 ISO 226: *Acoustics – normal equal-loudness contours* (International Organization for Standardization, Geneva 2003)

13.54 S.S. Stevens: On the psychophysical law, Psych. Rev. **64**, 153–181 (1957)

13.55 R.P. Hellman, J.J. Zwislocki: Some factors affecting the estimation of loudness, J. Acoust. Soc. Am. **35**, 687–694 (1961)

13.56 E.M. Relkin, J.R. Doucet: Is loudness simply proportional to the auditory nerve spike count?, J. Acoust. Soc. Am. **191**, 2735–2740 (1997)

13.57 H. Fletcher, W.A. Munson: Relation between loudness and masking, J. Acoust. Soc. Am. **9**, 1–10 (1937)

13.58 E. Zwicker: Über psychologische und methodische Grundlagen der Lautheit, Acustica **8**, 237–258 (1958)

13.59 E. Zwicker, B. Scharf: A model of loudness summation, Psych. Rev. **72**, 3–26 (1965)

13.60 B.R. Glasberg, B.C.J. Moore: A model of loudness applicable to time-varying sounds, J. Audio Eng. Soc. **50**, 331–342 (2002)

13.61 H. Fastl: Loudness evaluation by subjects and by a loudness meter. In: *Sensory Research – Multimodal Perspectives*, ed. by R.T. Verrillo (Erlbaum, Hillsdale, New Jersey 1993)

13.62 ANSI: *ANSI S3.4-2005. Procedure for the Computation of Loudness of Steady Sounds* (American National Standards Institute, New York 2005)

13.63 E. Zwicker, G. Flottorp, S.S. Stevens: Critical bandwidth in loudness summation, J. Acoust. Soc. Am. **29**, 548–557 (1957)

13.64 B. Scharf: Complex sounds and critical bands, Psychol. Bull. **58**, 205–217 (1961)

13.65 B. Scharf: Critical bands. In: *Foundations of Modern Auditory Theory*, ed. by J.V. Tobias (Academic, New York 1970)

13.66 B.C.J. Moore, B.R. Glasberg: The role of frequency selectivity in the perception of loudness, pitch and time. In: *Frequency Selectivity in Hearing*, ed. by B.C.J. Moore (Academic, London 1986)

13.67 G.A. Miller: Sensitivity to changes in the intensity of white noise and its relation to masking and loudness, J. Acoust. Soc. Am. **191**, 609–619 (1947)

13.68 R.R. Riesz: Differential intensity sensitivity of the ear for pure tones, Phys. Rev. **31**, 867–875 (1928)

13.69 N.F. Viemeister, S.P. Bacon: Intensity discrimination, increment detection, and magnitude estimation for 1-kHz tones, J. Acoust. Soc. Am. **84**, 172–178 (1988)

13.70 S.P. Bacon, N.F. Viemeister: Temporal modulation transfer functions in normal-hearing and hearing-impaired subjects, Audiology **24**, 117–134 (1985)

13.71 N.F. Viemeister, C.J. Plack: Time analysis. In: *Human Psychophysics*, ed. by W.A. Yost, A.N. Popper, R.R. Fay (Springer, New York 1993)

13.72 R. Plomp: The rate of decay of auditory sensation, J. Acoust. Soc. Am. **36**, 277–282 (1964)

13.73 M.J. Penner: Detection of temporal gaps in noise as a measure of the decay of auditory sensation, J. Acoust. Soc. Am. **61**, 552–557 (1977)

13.74 D. Ronken: Monaural detection of a phase difference between clicks, J. Acoust. Soc. Am. **47**, 1091–1099 (1970)

13.75 D.M. Green: Temporal acuity as a function of frequency, J. Acoust. Soc. Am. **54**, 373–379 (1973)

13.76 B.C.J. Moore, R.W. Peters, B.R. Glasberg: Detection of temporal gaps in sinusoids: Effects of frequency and level, J. Acoust. Soc. Am. **93**, 1563–1570 (1993)

13.77 A. Kohlrausch, R. Fassel, T. Dau: The influence of carrier level and frequency on modulation and beat-detection thresholds for sinusoidal carriers, J. Acoust. Soc. Am. **108**, 723–734 (2000)

13.78 B.C.J. Moore, B.R. Glasberg: Temporal modulation transfer functions obtained using sinusoidal carriers with normally hearing and hearing-impaired listeners, J. Acoust. Soc. Am. **110**, 1067–1073 (2001)

13.79 H. Fleischer: Modulationsschwellen von Schmalbandrauschen, Acustica **51**, 154–161 (1982)

13.80 T. Dau, B. Kollmeier, A. Kohlrausch: Modeling auditory processing of amplitude modulation: I. Detection and masking with narrowband carriers, J. Acoust. Soc. Am. **102**, 2892–2905 (1997)

13.81 T. Dau, B. Kollmeier, A. Kohlrausch: Modeling auditory processing of amplitude modulation: II. Spectral and temporal integration, J. Acoust. Soc. Am. **102**, 2906–2919 (1997)

13.82 T. Dau, J.L. Verhey, A. Kohlrausch: Intrinsic envelope fluctuations and modulation-detection thresholds for narrow-band noise carriers, J. Acoust. Soc. Am. **106**, 2752–2760 (1999)

13.83 N.F. Viemeister: Temporal modulation transfer functions based on modulation thresholds, J. Acoust. Soc. Am. **66**, 1364–1380 (1979)

13.84 B.C.J. Moore, B.R. Glasberg, C.J. Plack, A.K. Biswas: The shape of the ear's temporal window, J. Acoust. Soc. Am. **83**, 1102–1116 (1988)

13.85 R.H. Kay: Hearing of modulation in sounds, Physiol. Rev. **62**, 894–975 (1982)

13.86 T. Houtgast: Frequency selectivity in amplitude-modulation detection, J. Acoust. Soc. Am. **85**, 1676–1680 (1989)

13.87 S.P. Bacon, D.W. Grantham: Modulation masking: effects of modulation frequency, depth and phase, J. Acoust. Soc. Am. **85**, 2575–2580 (1989)

13.88 S.D. Ewert, T. Dau: Characterizing frequency selectivity for envelope fluctuations, J. Acoust. Soc. Am. **108**, 1181–1196 (2000)

13.89 C. Lorenzi, C. Soares, T. Vonner: Second-order temporal modulation transfer functions, J. Acoust. Soc. Am. **110**, 1030–1038 (2001)

13.90 A. Sek, B.C.J. Moore: Testing the concept of a modulation filter bank: The audibility of component modulation and detection of phase change in three-component modulators, J. Acoust. Soc. Am. **113**, 2801–2811 (2003)

13.91 C.D. Creelman: Human discrimination of auditory duration, J. Acoust. Soc. Am. **34**, 582–593 (1962)

13.92 S.M. Abel: Duration discrimination of noise and tone bursts, J. Acoust. Soc. Am. **51**, 1219–1223 (1972)

13.93 S.M. Abel: Discrimination of temporal gaps, J. Acoust. Soc. Am. **52**, 519–524 (1972)

13.94 P.L. Divenyi, W.F. Danner: Discrimination of time intervals marked by brief acoustic pulses of various intensities and spectra, Percept. Psychophys. **21**, 125–142 (1977)

13.95 J.H. Patterson, D.M. Green: Discrimination of transient signals having identical energy spectra, J. Acoust. Soc. Am. **48**, 894–905 (1970)

13.96 J. Zera, D.M. Green: Detecting temporal onset and offset asynchrony in multicomponent complexes, J. Acoust. Soc. Am. **93**, 1038–1052 (1993)

13.97 J.C. Risset, D.L. Wessel: Exploration of timbre by analysis and synthesis. In: *The Psychology of Music*, 2nd edn., ed. by D. Deutsch (Academic, San Diego 1999)

13.98 G. von Békésy: *Experiments in Hearing* (McGraw–Hill, New York 1960)

13.99 J.F. Schouten: The residue and the mechanism of hearing, Proc. Kon. Ned. Akad. Wetenschap. **43**, 991–999 (1940)

13.100 W.M. Siebert: Frequency discrimination in the auditory system: place or periodicity mechanisms, Proc. IEEE **58**, 723–730 (1970)

13.101 E. Zwicker: Masking and psychological excitation as consequences of the ear's frequency analysis. In: *Frequency Analysis and Periodicity Detection in Hearing*, ed. by R. Plomp, G.F. Smoorenburg (Sijthoff, Leiden 1970)

13.102 C.C. Wier, W. Jesteadt, D.M. Green: Frequency discrimination as a function of frequency and sensation level, J. Acoust. Soc. Am. **61**, 178–184 (1977)

13.103 A. Sek, B.C.J. Moore: Frequency discrimination as a function of frequency, measured in several ways, J. Acoust. Soc. Am. **97**, 2479–2486 (1995)

13.104 B.C.J. Moore: Relation between the critical bandwidth and the frequency-difference limen, J. Acoust. Soc. Am. **55**, 359 (1974)

13.105 J.L. Goldstein, P. Srulovicz: Auditory-nerve spike intervals as an adequate basis for aural frequency measurement. In: *Psychophysics and Physiology of Hearing*, ed. by E.F. Evans, J.P. Wilson (Academic, London 1977)

13.106 B.C.J. Moore, A. Sek: Detection of frequency modulation at low modulation rates: Evidence for a mechanism based on phase locking, J. Acoust. Soc. Am. **100**, 2320–2331 (1996)

13.107 G. Revesz: *Zur Grundlegung der Tonpsychologie* (Veit, Leipzig 1913)

13.108 A. Bachem: Tone height and tone chroma as two different pitch qualities, Acta Psych. **7**, 80–88 (1950)

13.109 W.D. Ward: Subjective musical pitch, J. Acoust. Soc. Am. **26**, 369–380 (1954)

13.110 F. Attneave, R.K. Olson: Pitch as a medium: A new approach to psychophysical scaling, Am. J. Psychol. **84**, 147–166 (1971)

13.111 K. Ohgushi, T. Hatoh: Perception of the musical pitch of high frequency tones. In: *Ninth International Symposium on Hearing: Auditory Physiology and Perception*, ed. by Y. Cazals, L. Demany, K. Horner (Pergamon, Oxford 1991)

13.112 S.S. Stevens: The relation of pitch to intensity, J. Acoust. Soc. Am. **6**, 150–154 (1935)

13.113 J. Verschuure, A.A. van Meeteren: The effect of intensity on pitch, Acustica **32**, 33–44 (1975)

13.114 G.S. Ohm: Über die Definition des Tones, nebst daran geknüpfter Theorie der Sirene und ähnlicher tonbildender Vorrichtungen, Annalen der Physik und Chemie **59**, 513–565 (1843)

13.115 J.F. Schouten: The residue revisited. In: *Frequency Analysis and Periodicity Detection in Hearing*, ed. by R. Plomp, G.F. Smoorenburg (Sijthoff, Leiden, The Netherlands 1970)

13.116 J.C.R. Licklider: Auditory frequency analysis. In: *Information Theorie*, ed. by C. Cherry (Academic, New York 1956)

13.117 E. de Boer: *On the 'residue' in hearing, Ph.D. Thesis* (University of Amsterdam, Amsterdam 1956)

13.118 J.L. Goldstein: An optimum processor theory for the central formation of the pitch of complex tones, J. Acoust. Soc. Am. **54**, 1496–1516 (1973)

13.119 E. Terhardt: Pitch, consonance, and harmony, J. Acoust. Soc. Am. **55**, 1061–1069 (1974)

13.120 G.F. Smoorenburg: Pitch perception of two-frequency stimuli, J. Acoust. Soc. Am. **48**, 924–941 (1970)

13.121 A.J.M. Houtsma, J.F.M. Fleuren: Analytic and synthetic pitch of two-tone complexes, J. Acoust. Soc. Am. **90**, 1674–1676 (1991)

13.122 J.L. Flanagan, M.G. Saslow: Pitch discrimination for synthetic vowels, J. Acoust. Soc. Am. **30**, 435–442 (1958)

13.123 B.C.J. Moore, B.R. Glasberg, M.J. Shailer: Frequency and intensity difference limens for harmonics within complex tones, J. Acoust. Soc. Am. **75**, 550–561 (1984)

13.124 B.C.J. Moore, B.R. Glasberg, R.W. Peters: Relative dominance of individual partials in determining the pitch of complex tones, J. Acoust. Soc. Am. **77**, 1853–1860 (1985)

13.125 R. Plomp: Pitch of complex tones, J. Acoust. Soc. Am. **41**, 1526–1533 (1967)

13.126 R.J. Ritsma: Frequencies dominant in the perception of the pitch of complex sounds, J. Acoust. Soc. Am. **42**, 191–198 (1967)

13.127 R.P. Carlyon, T.M. Shackleton: Comparing the fundamental frequencies of resolved and unresolved harmonics: Evidence for two pitch mechanisms?, J. Acoust. Soc. Am. **95**, 3541–3554 (1994)

13.128 A. Hoekstra, R.J. Ritsma: Perceptive hearing loss and frequency selectivity. In: *Psychophysics and Physiology of Hearing*, ed. by E.F. Evans, J.P. Wilson (Academic, London, England 1977)

13.129 A.J.M. Houtsma, J. Smurzynski: Pitch identification and discrimination for complex tones with many harmonics, J. Acoust. Soc. Am. **87**, 304–310 (1990)

13.130 B.C.J. Moore, S.M. Rosen: Tune recognition with reduced pitch and interval information, Q. J. Exp. Psychol. **31**, 229–240 (1979)

13.131 R. Meddis, M. Hewitt: A computational model of low pitch judgement. In: *Basic Issues in Hearing*, ed. by H. Duifhuis, J.W. Horst, H.P. Wit (Academic, London 1988)

13.132 R. Meddis, M. Hewitt: Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification, J. Acoust. Soc. Am. **89**, 2866–2882 (1991)

13.133 B.C.J. Moore: *An Introduction to the Psychology of Hearing*, 2nd edn. (Academic, London 1982)

13.134 B.C.J. Moore: *An Introduction to the Psychology of Hearing*, 5th edn. (Academic, San Diego 2003)

13.135 P. Srulovicz, J.L. Goldstein: A central spectrum model: a synthesis of auditory-nerve timing and place cues in monaural communication of frequency spectrum, J. Acoust. Soc. Am. **73**, 1266–1276 (1983)

13.136 R. Plomp: Timbre as a multidimensional attribute of complex tones. In: *Frequency Analysis and Periodicity Detection in Hearing*, ed. by R. Plomp, G.F. Smoorenburg (Sijthoff, Leiden 1970)

13.137 G. von Bismarck: Sharpness as an attribute of the timbre of steady sounds, Acustica **30**, 159–172 (1974)

13.138 R. Plomp: *Aspects of Tone Sensation* (Academic, London 1976)

13.139 R.D. Patterson: A pulse ribbon model of monaural phase perception, J. Acoust. Soc. Am. **82**, 1560–1586 (1987)

13.140 R. Plomp, H.J.M. Steeneken: Effect of phase on the timbre of complex tones, J. Acoust. Soc. Am. **46**, 409–421 (1969)

13.141 A.J. Watkins: Central, auditory mechanisms of perceptual compensation for spectral-envelope distortion, J. Acoust. Soc. Am. **90**, 2942–2955 (1991)

13.142 J.F. Schouten: The perception of timbre, 6th International Conference on Acoustics 1, GP-6-2 (1968)

13.143 R.D. Patterson: The sound of a sinusoid: Spectral models, J. Acoust. Soc. Am. **96**, 1409–1418 (1994)

13.144 R.D. Patterson: The sound of a sinusoid: Time-interval models, J. Acoust. Soc. Am. **96**, 1419–1428 (1994)

13.145 M.A. Akeroyd, R.D. Patterson: Discrimination of wideband noises modulated by a temporally asymmetric function, J. Acoust. Soc. Am. **98**, 2466–2474 (1995)

13.146 H.F. Pollard, E.V. Jansson: A tristimulus method for the specification of musical timbre, Acustica **51**, 162–171 (1982)

13.147 S. Handel: Timbre perception and auditory object identification. In: *Hearing*, ed. by B.C.J. Moore (Academic, San Diego 1995)

13.148 A.W. Mills: On the minimum audible angle, J. Acoust. Soc. Am. **30**, 237–246 (1958)

13.149 L. Rayleigh: On our perception of sound direction, Phil. Mag. **13**, 214–232 (1907)

13.150 E.R. Hafter: Spatial hearing and the duplex theory: How viable?. In: *Dynamic Aspects of Neocortical Function*, ed. by G.M. Edelman, W.E. Gall, W.M. Cowan (Wiley, New York 1984)

13.151 G.B. Henning: Detectability of interaural delay in high-frequency complex waveforms, J. Acoust. Soc. Am. **55**, 84–90 (1974)

13.152 D.W. Batteau: The role of the pinna in human localization, Proc. Roy. Soc. B. **168**, 158–180 (1967)

13.153 J. Blauert: *Spatial Hearing: The Psychophysics of Human Sound Localization* (MIT Press, Cambridge, Mass 1997)

13.154 W.M. Hartmann, A. Wittenberg: On the externalization of sound images, J. Acoust. Soc. Am. **99**, 3678–3688 (1996)

13.155 H. Haas: Über den Einfluss eines Einfachechos an die Hörsamkeit von Sprache, Acustica **1**, 49–58 (1951)

13.156 H. Wallach, E.B. Newman, M.R. Rosenzweig: The precedence effect in sound localization, Am. J. Psychol. **62**, 315–336 (1949)

13.157 R.Y. Litovsky, H.S. Colburn, W.A. Yost, S.J. Guzman: The precedence effect, J. Acoust. Soc. Am. **106**, 1633–1654 (1999)

13.158 A.S. Bregman: *Auditory Scene Analysis: The Perceptual Organization of Sound* (Bradford Books, MIT Press, Cambridge, Mass. 1990)

13.159 A.S. Bregman, S. Pinker: Auditory streaming and the building of timbre, Canad. J. Psychol. **32**, 19–31 (1978)

13.160 C.J. Darwin, R.P. Carlyon: Auditory grouping. In: *Hearing*, ed. by B.C.J. Moore (Academic, San Diego 1995)

13.161 D.E. Broadbent, P. Ladefoged: On the fusion of sounds reaching different sense organs, J. Acoust. Soc. Am. **29**, 708–710 (1957)

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

13.163 P.F. Assmann, A.Q. Summerfield: Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies, J. Acoust. Soc. Am. **88**, 680–697 (1990)

13.164 J.D. McKeown, R.D. Patterson: The time course of auditory segregation: Concurrent vowels that vary in duration, J. Acoust. Soc. Am. **98**, 1866–1877 (1995)

13.165 B.C.J. Moore, B.R. Glasberg, R.W. Peters: Thresholds for hearing mistuned partials as separate tones in harmonic complexes, J. Acoust. Soc. Am. **80**, 479–483 (1986)

13.166 B. Roberts, J.M. Brunstrom: Perceptual segregation and pitch shifts of mistuned components in harmonic complexes and in regular inharmonic complexes, J. Acoust. Soc. Am. **104**, 2326–2338 (1998)

13.167 B. Roberts, J.M. Brunstrom: Perceptual fusion and fragmentation of complex tones made inharmonic by applying different degrees of frequency shift and spectral stretch, J. Acoust. Soc. Am. **110**, 2479–2490 (2001)

13.168 R. Meddis, M. Hewitt: Modeling the identification of concurrent vowels with different fundamental frequencies, J. Acoust. Soc. Am. **91**, 233–245 (1992)

13.169 A. de Cheveigné, S. McAdams, C.M.H. Marin: Concurrent vowel identification. II. Effects of phase, harmonicity and task, J. Acoust. Soc. Am. **101**, 2848–2856 (1997)

13.170 A. de Cheveigné, H. Kawahara, M. Tsuzaki, K. Aikawa: Concurrent vowel identification. I. Effects of relative amplitude and F0 difference, J. Acoust. Soc. Am. **101**, 2839–2847 (1997)

13.171 A. de Cheveigné: Concurrent vowel identification. III. A neural model of harmonic interference cancellation, J. Acoust. Soc. Am. **101**, 2857–2865 (1997)

13.172 R.A. Rasch: The perception of simultaneous notes such as in polyphonic music, Acustica **40**, 21–33 (1978)

13.173 C.J. Darwin, N.S. Sutherland: Grouping frequency components of vowels: when is a harmonic not a harmonic?, Q. J. Exp. Psychol. **36A**, 193–208 (1984)

13.174 B. Roberts, B.C.J. Moore: The influence of extraneous sounds on the perceptual estimation of first-formant frequency in vowels under conditions of asynchrony, J. Acoust. Soc. Am. **89**, 2922–2932 (1991)

13.175 E. Zwicker: 'Negative afterimage' in hearing, J. Acoust. Soc. Am. **36**, 2413–2415 (1964)

13.176 A.Q. Summerfield, A.S. Sidwell, T. Nelson: Auditory enhancement of changes in spectral amplitude, J. Acoust. Soc. Am. **81**, 700–708 (1987)

13.177 A.S. Bregman, J. Abramson, P. Doehring, C.J. Darwin: Spectral integration based on common amplitude modulation, Percept. Psychophys. **37**, 483–493 (1985)

13.178 J.W. Hall, J.H. Grose: Comodulation masking release and auditory grouping, J. Acoust. Soc. Am. **88**, 119–125 (1990)

13.179 B.C.J. Moore, M.J. Shailer: Comodulation masking release as a function of level, J. Acoust. Soc. Am. **90**, 829–835 (1991)

13.180 B.C.J. Moore, M.J. Shailer, M.J. Black: Dichotic interference effects in gap detection, J. Acoust. Soc. Am. **93**, 2130–2133 (1993)

13.181 Q. Summerfield, J.F. Culling: Auditory segregation of competing voices: absence of effects of FM or AM coherence, Phil. Trans. R. Soc. Lond. B **336**, 357–366 (1992)

13.182 M.F. Cohen, X. Chen: Dynamic frequency change among stimulus components: Effects of coherence on detectability, J. Acoust. Soc. Am. **92**, 766–772 (1992)

13.183 M.H. Chalikia, A.S. Bregman: The perceptual segregation of simultaneous vowels with harmonic, shifted, and random components, Percept. Psychophys. **53**, 125–133 (1993)

13.184 S. McAdams: Segregation of concurrent sounds. I.: Effects of frequency modulation coherence, J. Acoust. Soc. Am. **86**, 2148–2159 (1989)

13.185 R.P. Carlyon: Discriminating between coherent and incoherent frequency modulation of complex tones, J. Acoust. Soc. Am. **89**, 329–340 (1991)

13.186 R.P. Carlyon: Further evidence against an across-frequency mechanism specific to the detection of frequency modulation (FM) incoherence between resolved frequency components, J. Acoust. Soc. Am. **95**, 949–961 (1994)

13.187 J. Lyzenga, B.C.J. Moore: Effect of FM coherence for inharmonic stimuli: FM-phase discrimination and identification of artificial double vowels, J. Acoust. Soc. Am. **117**, 1314–1325 (2005)

13.188 C.M.H. Marin, S. McAdams: Segregation of concurrent sounds. II: Effects of spectral envelope tracing, frequency modulation coherence, and frequency modulation width, J. Acoust. Soc. Am. **89**, 341–351 (1991)

13.189 S. Furukawa, B.C.J. Moore: Across-frequency processes in frequency modulation detection, J. Acoust. Soc. Am. **100**, 2299–2312 (1996)

13.190 S. Furukawa, B.C.J. Moore: Dependence of frequency modulation detection on frequency modulation coherence across carriers: Effects of modulation rate, harmonicity and roving of the carrier frequencies, J. Acoust. Soc. Am. **101**, 1632–1643 (1997)

13.191 S. Furukawa, B.C.J. Moore: Effect of the relative phase of amplitude modulation on the detection of modulation on two carriers, J. Acoust. Soc. Am. **102**, 3657–3664 (1997)

13.192 R.P. Carlyon: Detecting coherent and incoherent frequency modulation, Hear. Res. **140**, 173–188 (2000)

13.193 M. Kubovy, J.E. Cutting, R.M. McGuire: Hearing with the third ear: dichotic perception of a melody without monaural familiarity cues, Science **186**, 272–274 (1974)

13.194 J.F. Culling: Auditory motion segregation: a limited analogy with vision, J. Exp. Psychol.: Human Percept. Perf. **26**, 1760–1769 (2000)

13.195 M.A. Akeroyd, B.C.J. Moore, G.A. Moore: Melody recognition using three types of dichotic-pitch stimulus, J. Acoust. Soc. Am. **110**, 1498–1504 (2001)

13.196 T.M. Shackleton, R. Meddis: The role of interaural time difference and fundamental frequency difference in the identification of concurrent vowel pairs, J. Acoust. Soc. Am. **91**, 3579–3581 (1992)

13.197 J.F. Culling, Q. Summerfield: Perceptual separation of concurrent speech sounds: Absence of across-frequency grouping by common interaural delay, J. Acoust. Soc. Am. **98**, 785–797 (1995)

13.198 C.J. Darwin, R.W. Hukin: Auditory objects of attention: the role of interaural time differences, J. Exp. Psychol.: Human Percept. Perf. **25**, 617–629 (1999)

13.199 G.A. Miller, G.A. Heise: The trill threshold, J. Acoust. Soc. Am. **22**, 637–638 (1950)

13.200 A.S. Bregman, J. Campbell: Primary auditory stream segregation and perception of order in rapid sequences of tones, J. Exp. Psychol. **89**, 244–249 (1971)

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

13.202 L.P.A.S. van Noorden: Rhythmic fission as a function of tone rate, IPO Annual Prog. Rep. **6**, 9–12 (1971)

13.203 A.S. Bregman, G. Dannenbring: The effect of continuity on auditory stream segregation, Percept. Psychophys. **13**, 308–312 (1973)

13.204 A.S. Bregman: Auditory streaming is cumulative, J. Exp. Psychol.: Human Percept. Perf. **4**, 380–387 (1978)

13.205 W.L. Rogers, A.S. Bregman: An experimental evaluation of three theories of auditory stream segregation, Percept. Psychophys. **53**, 179–189 (1993)

13.206 W.L. Rogers, A.S. Bregman: Cumulation of the tendency to segregate auditory streams: resetting by changes in location and loudness, Percept. Psychophys. **60**, 1216–1227 (1998)

13.207 B.C.J. Moore, H. Gockel: Factors influencing sequential stream segregation, Acta Acustica – Acustica **88**, 320–333 (2002)

13.208 W.M. Hartmann, D. Johnson: Stream segregation and peripheral channeling, Music Percept. **9**, 155–184 (1991)

13.209 P.G. Singh, A.S. Bregman: The influence of different timbre attributes on the perceptual segregation of complex-tone sequences, J. Acoust. Soc. Am. **102**, 1943–1952 (1997)

13.210 J. Vliegen, B.C.J. Moore, A.J. Oxenham: The role of spectral and periodicity cues in auditory stream segregation, measured using a temporal discrimination task, J. Acoust. Soc. Am. **106**, 938–945 (1999)

13.211 J. Vliegen, A.J. Oxenham: Sequential stream segregation in the absence of spectral cues, J. Acoust. Soc. Am. **105**, 339–346 (1999)

13.212 P. Iverson: Auditory stream segregation by musical timbre: effects of static and dynamic acoustic attributes, J. Exp. Psychol.: Human Percept. Perf. **21**, 751–763 (1995)

13.213 B. Roberts, B.R. Glasberg, B.C.J. Moore: Primitive stream segregation of tone sequences without differences in F0 or passband, J. Acoust. Soc. Am. **112**, 2074–2085 (2002)

13.214 H. Gockel, R.P. Carlyon, C. Micheyl: Context dependence of fundamental-frequency discrimination: Lateralized temporal fringes, J. Acoust. Soc. Am. **106**, 3553–3563 (1999)

13.215 W.J. Dowling: Rhythmic fission and perceptual organization, J. Acoust. Soc. Am. **44**, 369 (1968)

13.216 W.J. Dowling: The perception of interleaved melodies, Cognitive Psychol. **5**, 322–337 (1973)

13.217 D. Deutsch: Two-channel listening to musical scales, J. Acoust. Soc. Am. **57**, 1156–1160 (1975)

13.218 D.E. Broadbent, P. Ladefoged: Auditory perception of temporal order, J. Acoust. Soc. Am. **31**, 151–159 (1959)

13.219 R.M. Warren, C.J. Obusek, R.M. Farmer, R.P. Warren: Auditory sequence: confusion of patterns other than speech or music, Science N.Y. **164**, 586–587 (1969)

13.220 R.M. Warren: Auditory temporal discrimination by trained listeners, Cognitive Psychol. **6**, 237–256 (1974)

13.221 P.L. Divenyi, I.J. Hirsh: Identification of temporal order in three-tone sequences, J. Acoust. Soc. Am. **56**, 144–151 (1974)

13.222 R. Cusack, B. Roberts: Effects of differences in timbre on sequential grouping, Percept. Psychophys. **62**, 1112–1120 (2000)

13.223 K. Koffka: *Principles of Gestalt Psychology* (Harcourt and Brace, New York 1935)

13.224 C.J. Darwin, C.E. Bethell-Fox: Pitch continuity and speech source attribution, J. Exp. Psychol.: Hum. Perc. Perf. **3**, 665–672 (1977)

13.225 A.S. Bregman, A. Rudnicky: Auditory segregation: stream or streams?, J. Exp. Psychol.: Human Percept. Perf. **1**, 263–267 (1975)

13.226 A.S. Bregman: The meaning of duplex perception: sounds as transparent objects. In: *The Psychophysics of Speech Perception*, ed. by M.E.H. Schouten (Martinus Nijhoff, Dordrecht 1987)

13.227 T. Houtgast: Psychophysical evidence for lateral inhibition in hearing, J. Acoust. Soc. Am. **51**, 1885–1894 (1972)

13.228 W.R. Thurlow: An auditory figure-ground effect, Am. J. Psychol. **70**, 653–654 (1957)

13.229 R.M. Warren, C.J. Obusek, J.M. Ackroff: Auditory induction: perceptual synthesis of absent sounds, Science **176**, 1149–1151 (1972)

13.230 V. Ciocca, A.S. Bregman: Perceived continuity of gliding and steady-state tones through interrupting noise, Percept. Psychophys. **42**, 476–484 (1987)

13.231 G.A. Miller, J.C.R. Licklider: The intelligibility of interrupted speech, J. Acoust. Soc. Am. **22**, 167–173 (1950)

13.232 D. Dirks, D. Bower: Effects of forward and backward masking on speech intelligibility, J. Acoust. Soc. Am. **47**, 1003–1008 (1970)

13.233 R.M. Warren: Perceptual restoration of missing speech sounds, Science **167**, 392–393 (1970)