11

Further Analysis in Hearing

Chapter 9 presented two important concepts in psychoacoustic experimentation – masking and critical bands. Chapter 10 presented the four central quantities or dimensions of psychoacoustics – pitch, loudness, timbre, and subjective duration, all of which are relatively well defined and orthogonal to each other, except perhaps timbre. Timbre is a multidimensional and complex measure, and a number of quantities that describe the perception of timbre have been defined in the literature; these can be regarded as subcategories of timbre.

This chapter describes a few of these quantities which are useful in the research on psychoacoustics or in technical applications. These quantities are sharpness, roughness, fluctuation strength, tonality, consonance, and dissonance. These quantities also provide an interesting connection to the intervals in music scales.

Sound signals are primarily one-dimensional, time-dependent functions. However, our hearing distils them into features and characteristics that depend both on time and frequency, which makes sound a multidimensional entity for neural processing. The psychoacoustic foundations of our perception of sound and music, which is an art form that takes full advantage of the time–frequency structures of sound, are discussed in this chapter. In natural listening conditions, our hearing attempts to represent sound scenarios as objects, just as the other senses do. Certain complex sound scenarios, such as sound in environments containing many sources or music presented by an ensemble of instruments, are often structured into auditory streams in our hearing.

11.1 Sharpness

When subjects are asked 'how sharp is the sound?' on being presented with sounds having different spectral content, a relatively stable, repeatable, and subject-independent response can be measured. The resulting response is the psychoacoustic quantity called *sharpness* (Fastl and Zwicker, 2007, Chapter 9). The higher in frequency the centre of gravity of the amplitude spectrum is located, the higher is the sharpness measured. Figure 11.1 shows the sharpness of narrowband noise with a width of 1 Bark as a function of frequency. The figure also shows the sharpness for low-pass and high-pass filtered noise as a function of cut-off frequency.

Communication Acoustics: An Introduction to Speech, Audio, and Psychoacoustics, First Edition.
Ville Pulkki and Matti Karjalainen.

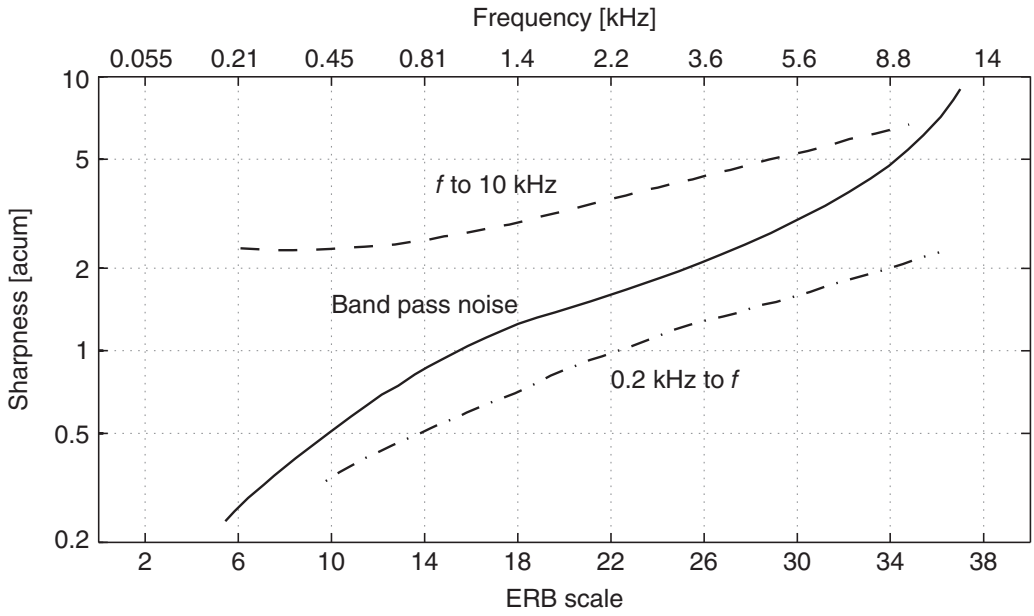

Figure 11.1 The sharpness of narrowband noise (solid line), high-pass filtered noise (upper cutoff is at 10 kHz), and low-pass filtered noise (lower cutoff is at 200 Hz) as a function of the centre frequency or cutoff frequency. Adapted from Fastl and Zwicker (2007).

The unit of sharpness, the *acum*, is defined such that narrowband noise of width 1 Bark and centre frequency 1 kHz at an SPL of 60 dB produces a sharpness value of 1 acum. Psychoacoustic testing reveals that the level of sound also affects the perceived sharpness, but not very prominently. Increasing the level from 30 dB to 90 dB only doubles the sharpness value (Fastl and Zwicker, 2007, Chapter 9).

For narrowband noise, the sharpness rises monotonically on the frequency scale. Below 1 kHz and above 4 kHz the rise is steeper. The steady rise of sharpness S in the middle frequencies and the steep rise at high frequencies can be modelled simply by defining the gain $g(z)$ on the Bark scale z, as in Figure 11.2, and by writing

$$S \sim g(z)\, z. \tag{11.1}$$

Figure 11.1 shows that the sharpness of broadband sounds depends on where the 'centre of gravity' of specific loudness is located, and when it is situated at higher frequencies, particularly high sharpness is obtained. This observation can be used to derive a simple computational model for sharpness, which can be written as

$$S = 0.11\, \frac{\int_0^{24\,\mathrm{Bark}} N'(z)\, g(z)\, z\, \mathrm{d}z}{\int_0^{24\,\mathrm{Bark}} N'(z) \mathrm{d}z}, \tag{11.2}$$

where $N'(z)$ is the specific loudness. The equation does not represent the relationship between the level of the sound and sharpness. Although the equation is relatively simple, it provides a fairly accurate estimation of sharpness for different sounds.

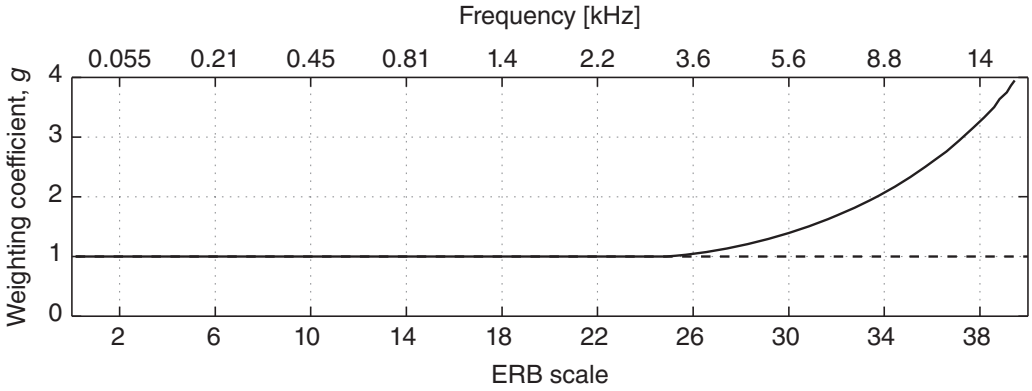

Figure 11.2 The gain factor $g(z)$ as a function of the Bark scale z used to compute sharpness. Adapted from Fastl and Zwicker (2007).

11.2 Detection of Modulation and Sound Onset

Considerable changes in the acoustic characteristics of sound with time are perceived as beating, warbling, impulsiveness, fluctuation, or just changing in general. If the level of sound fluctuates relatively slowly over a certain range of rates, between about 1 Hz and 16 Hz, a significant *fluctuation strength* may be perceived. When the modulation rate far exceeds 16 Hz, our hearing is unable to follow the level of sound, and the sound is perceived as *rough*, associated with the psychoacoustic quantity *roughness*. A fast-rising sound is perceived as *impulsive*, and the corresponding psychoacoustic quantity for this perception is *impulsiveness*.

11.2.1 Fluctuation Strength

The simplest examples of fluctuation in sound are amplitude and frequency modulation. Such modulations give rise to the perception of the psychoacoustic quantity called *fluctuation strength* (Fastl and Zwicker, 2007). If the modulation has a relatively low frequency, below about 0.5 Hz, our hearing does not detect it anymore, since the short-term auditory memory does not have a very good reference to the past. If the frequency of modulation is higher than about 16 Hz, the sluggishness of the hearing mechanisms starts to limit the resolution of fluctuation, and the modulations are perceived to produce *roughness*, as discussed later in Section 11.3. The hearing is most sensitive to fluctuations and modulation near the frequency 4 Hz.

The unit of fluctuation strength is the *vacil*, where the reference point is selected by prescribing that a 1-kHz sinusoid at a level of 60 dB with 100% amplitude modulation at 4 Hz produces a fluctuation strength of magnitude 1 vacil. The fluctuation strength is shown in Figure 11.3 as a function of modulation frequency for three cases: amplitude-modulated broadband noise, an amplitude-modulated tone, and a frequency-modulated tone. The curves are similar, and their peak is found at about 4 Hz (Fastl and Zwicker, 2007, Chapter 10).

The dependence of fluctuation strength on modulation depth is shown in Figure 11.4 (AM BBN) for broadband noise, where it can be seen that modulation of a few dB is needed to produce notable fluctuation strength, and that the phenomenon saturates at a modulation depth of 20–30 dB. The corresponding dependence of a frequency-modulated tone on frequency deviation is shown in the same figure in the panel labelled FM tone. Increasing the frequency deviation is seen to cause increasing fluctuation strength, and no saturation effects

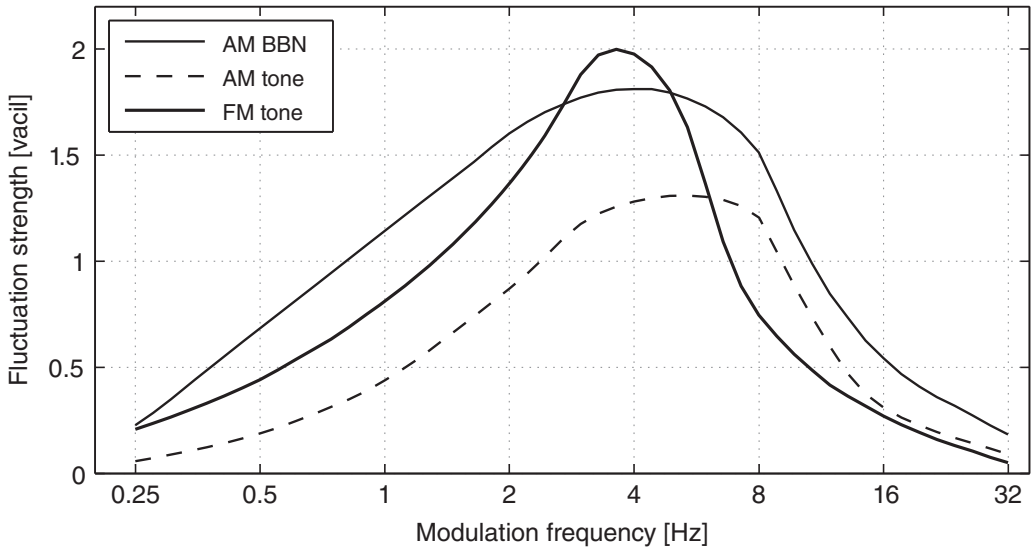

Figure 11.3 Fluctuation strength as a function of modulation frequency for three cases: (AM BBN) amplitude-modulated white noise with 40-dB modulation depth, (AM tone) amplitude-modulated tone with 40-dB modulation depth, and (FM tone) frequency-modulated 1-kHz tone with ± 700 Hz frequency variation. Adapted from Fastl and Zwicker (2007).

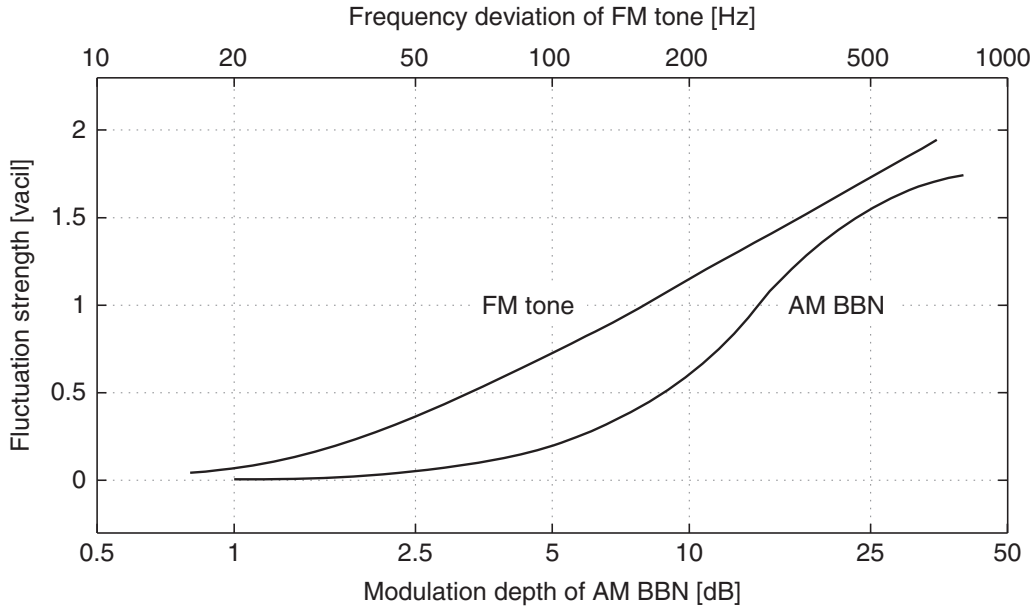

Figure 11.4 The fluctuation strength of a 4-Hz amplitude-modulated broadband noise as a function of modulation depth in dB (AM BBN). The fluctuation strength of a 1.5-kHz tone at a level of 70 dB and frequency-modulated at the rate of 4 Hz as a function of the modulation deviation in Hz (FM tone). Adapted from Fastl and Zwicker (2007).

are seen. The fluctuation strength also depends on loudness. For example, a change in the level of amplitude-modulated sound from 50 dB to 90 dB causes a fivefold increase in fluctuation strength.

The information presented so far suggests that our hearing forms the fluctuation strength percept by summing the modulation in the level of the signal in all of the auditory bands. The

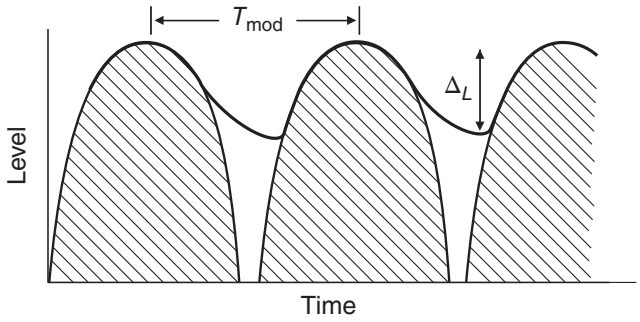

Figure 11.5 The temporal masking pattern caused by a sinusoidally amplitude-modulated masker, and the depth of the pattern ΔL and modulation period $T_{\mathrm{mod}} = 1/f_{\mathrm{mod}}$. Adapted from Fastl and Zwicker (2007).

principles illustrated in Figure 10.14 on page 186 can also be applied in this context, since they also explain the formation of the fluctuation strength with both amplitude and frequency modulation. Although the modulated signal components reside only in a narrow frequency region, the spreading function in the basilar membrane causes the effect of modulation to be present over a much wider region.

A schematic diagram of the phenomenon leading to the perception of fluctuation strength is shown in Figure 11.5. In the illustration, the input signal is an amplitude-modulated sinusoid with a modulation frequency of f_{mod}. The level of excitation in a single auditory band is shown as the hatched area, which exhibits strong modulation with time. The thick line shows the specific loudness schematically, which does not follow fast modulation. The quantity ΔL represents the depth of modulation of the temporal masking pattern, and the fluctuation strength F can be estimated as

$$F \sim \frac{\Delta L}{(f_{\mathrm{mod}}/4\,[\mathrm{Hz}]) + (4\,[\mathrm{Hz}]/f_{\mathrm{mod}})} \tag{11.3}$$

A slightly more general dependence between amplitude- and frequency-modulated signals can be written (Fastl and Zwicker, 2007, Chapter 10), if $\Delta L(z)$ is known:

$$F[\mathrm{vacil}] = \frac{0.008 \int_0^{24\,\mathrm{Bark}} (\Delta L(z))\mathrm{d}z}{(f_{\mathrm{mod}}/4[\mathrm{Hz}]) + (4[\mathrm{Hz}]/f_{\mathrm{mod}})}. \tag{11.4}$$

This relation can be implemented if $\Delta L(z)$ is measured with a suitable time-dependent auditory model. The denominator may also be implemented, if f_{mod} is known, or is measured by band-pass filtering L.

11.2.2 Impulsiveness

A sound is perceived to be 'impulsive', if it contains transients where the level of sound increases rapidly. As shown in Figure 3.13, impulses produce a very strong response in the auditory nerve, while on the other hand, as shown in Figure 10.18, very short sound events produce lower responses than their level would imply. However, transients are often perceived as an increase in annoyance caused by noise, and they very effectively grab a subject's attention. This attention steering may be seen as a method to make one aware of one's surroundings, since it may carry valid information about active processes that are often caused by a change

of state in physical objects. A strongly impulsive sound also contains high frequencies, since the sound cannot be very short without them.

There is no simple psychoacoustic measure for the *impulsiveness* of sound. In noise measurements there are some measures used, where the levels of measured noise are characterized as long-term mean levels. For example, the sound pressure level can be measured so that it exceeds the average 10% of the time. When the measured level is compared to the equivalent level of the noise, an estimate of impulsiveness is obtained. It has also been proposed that onsets which define impulsiveness can be identified by finding the regions where the positive slope of the instantaneous sound pressure level exceeds 10 dB/s (Pedersen, 2001).

11.3 Roughness

When two tones close to each other in frequency are added, the resulting broadband signal contains amplitude fluctuation at frequencies corresponding to the difference in the frequencies. This fluctuation is called *beating*. The resulting perception of the auditory event is also interesting, as it depends relatively systematically on the difference in frequency. This is shown conceptually in Figure 11.6.

When Δf is small, a slight and often pleasant beating is perceived, which can be measured as a fluctuation in the strength of sound. When Δf exceeds 10–15 Hz, the percept turns more 'rough', which is generally perceived to be unpleasant. After a certain value of Δf, the tones are not anymore fused, but instead are resolved and their frequencies are perceived independently of each other, however, with some roughness still present. When Δf is increased further, and the width of a critical band is exceeded, the roughness vanishes, and the presence of the two tones no longer has any influence on perception.

Roughness (Daniel and Weber, 1997; Fastl and Zwicker, 2007) is a psychoacoustic quantity which is caused by relatively fast amplitude modulations (15–300 Hz) that take place for stimuli within the range of a critical band. For example, narrowband noise always sounds slightly rough, since there are random amplitude fluctuations in it.

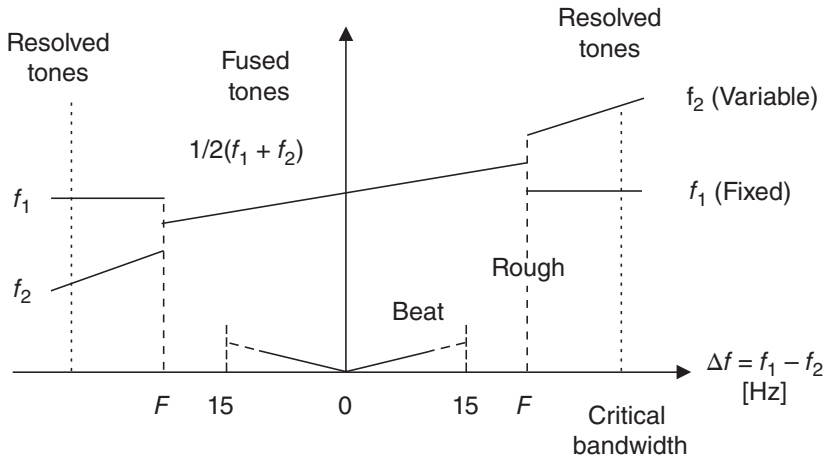

Figure 11.6 The nature of auditory perception of two sinusoids, one with a fixed frequency f_1 and another with a variable frequency f_2. The frequency difference is shown on the abscissa and the ordinate shows the frequency corresponding to the fused or resolved tones. Adapted from Roederer (1975) with permission from Academic Press.

The unit of roughness is the *asper*, and 1 asper is obtained with a 1-kHz tone which is 100% amplitude-modulated at a rate of 70 Hz. The roughness of such a sound is presented in Figure 11.7 for different modulation depths (Fastl and Zwicker, 2007, Chapter 11). Figure 11.8 presents roughness as a function of modulation frequency for different carrier frequencies. Most of the curves shown in Figure 11.8 have a coinciding low-pass shape with their maximum near 70 Hz, but the results with low-frequency sinusoids are different. At low frequencies, the width of the critical band (in hertz) is narrower than at high frequencies, and the maximum roughness is obtained with lower values of modulation frequency than at higher frequencies. Figure 11.8 shows that roughness can be explained as a kind of band-pass operation. The frequency of amplitude modulation present in the critical band can be thought to be band-pass filtered by a filter with the frequency response shown in Figure 11.8, and the level of the output of the filter is related to the value of roughness.

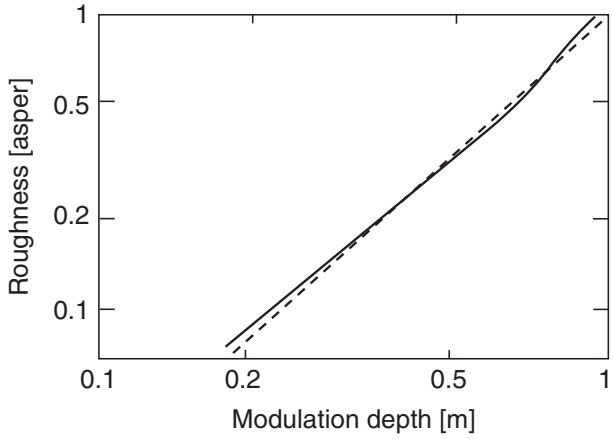

Figure 11.7 The solid line presents the roughness of an amplitude-modulated tone as a function of modulation depth, when the frequency of the tone is 1 kHz and the modulation frequency is 70 Hz. The dashed line is a linear fit of the roughness for reference. Fastl and Zwicke 2007.

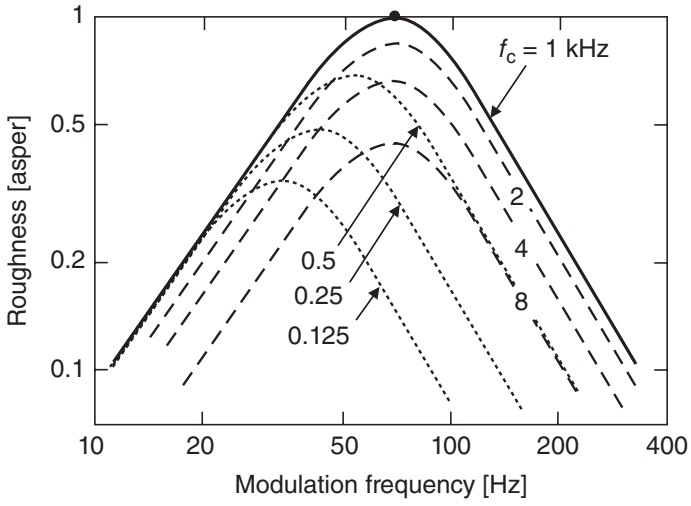

Figure 11.8 The roughness of a 100%-amplitude-modulated tone as a function of the modulation frequency. The different curves show the roughness for different carrier frequencies. Fastl and Zwicke 2007.

Roughness is also perceived to be high when broadband noise is amplitude modulated. Also, when a sinusoid is frequency modulated, the auditory event is perceived to be rough. In both cases the signal produces roughness in many critical bands. In auditory modelling, the roughness perception caused by an arbitrary sound can be estimated by summing the partial roughness created in each critical band.

11.4 Tonality

Tonality (or *tonalness*) is similar to the voicing character of speech: tonality is low with noisy sounds and high with sounds containing prominent frequency components, or tones. The tonality of a sound also increases with one or many high-amplitude and narrowband components, where the listener resolves either the fundamental frequency or a tonal component of the sound. Tonality is also related to pitch salience. When pitch salience is high, such as with pure tones or with narrowband noises, the tonality is high as well. However, with some signals, such as bell sounds, pitch is ambiguous, producing low pitch salience but yet high tonality. Note that tonality in psychoacoustics is a quantity which reflects how much the sound contains the characteristics of a *tone*. It is not to be confused with the terms from music, *tonality*, *tonal*, *atonal*, and *microtonal*, which are descriptors of musical scales and styles of musical composition.

There is no really clear method to measure tonality, although there are several starting points from which it can be quantified. Some methods are described below. A simple measure for tonality is the *tone-to-noise ratio, T/N*, which has also been standardized in ISO-7779 (2010). It measures the level of a tone compared to the level of the surrounding frequency band. Unfortunately, if there are multiple tones present in a critical band, the T/N method does not match the perceived tonality. The *prominence ratio, PR* (Bienvenue and Nobile, 1991) is a better approach in such cases; here, the power of the critical band containing the prominent tone is compared to the mean power in adjacent critical bands.

Tonality is also discussed and modelled by (Terhardt *et al.*, 1982, 1996). In the model, the local maxima are sought from the spectra. If the level difference between the maximum and its surroundings is more than 7 dB, the signal component corresponding to the maximum is thought to be tonal.

Tonality has also been studied in the context of audio technologies and speech transmission, since knowing how well a partial of a harmonic or noise masks another partial or noise is relevant in audio coding (Johnston, 1985). It has been found that noise is masked easier by a partial than vice versa. The *spectral flatness measure* (SFM) is first computed as the ratio of the geometric mean G_m to the arithmetic mean A_m of the power spectrum of the signal in dB,

$$\mathrm{SFM_{dB}} = 10\,\log_{10}\frac{G_\mathrm{m}}{A_\mathrm{m}}, \tag{11.5}$$

and the tonality α is then computed as

$$\alpha = \min\left(\frac{\mathrm{SFM_{dB}}}{\mathrm{SFM_{dBmax}}}, 1\right), \tag{11.6}$$

where $\mathrm{SFM_{dBmax}} = -60\,\mathrm{dB}$ is a reference value. If $\mathrm{SFM_{dB}} = 0\,\mathrm{dB}$, the signal is set to have a tonality value of 0 and thus to have a noise-like character. If $\mathrm{SFM_{dB}} < -60\,\mathrm{dB}$, the tonality is 1, and it is considered to have a tone-like character. Johnston (1985) also explains how the estimated tonality α is further used to estimate the audibility of some component of sound, so that it is retained or discarded during audio coding.

11.5 Discrimination of Changes in Signal Magnitude and Phase Spectra

The previous and the present chapters have discussed many properties of hearing, starting from simple signals and considering only their amplitude spectra. The analysis of basic psychoacoustic quantities with such signals helps to explain many phenomena in hearing. On the other hand, audio and speech signals can also be analysed by objective means, as shown in Chapter 3, for example by using spectral analysis, time–frequency plots, and spectrograms. This section focuses on the characteristics of hearing in relation to signal analysis, especially the ability to hear different characteristics in the magnitude and phase spectra. This helps develop an understanding of quality evaluation of audio techniques, which will be discussed further in Chapter 17.

11.5.1 Adaptation to the Magnitude Spectrum

A remarkable property of hearing is its ability to adapt to the acoustics of different listening conditions. To help in identifying a sound event across a transmission channel, it appears that listeners try to remove spectral distortion caused by the channel. Listeners seem to compensate for, or adapt to, the channel (Olive *et al.*, 1995; Pike *et al.*, 2013). In other words, it seems that listeners are able, at least partially, to inverse-filter the spectral effect of the channel to more accurately estimate the signal emitted by the source (Toole, 2006; Watkins, 1991).

The ability to do this is most probably a result of neural processes within the hearing system. For example, a study by Summerfield *et al.* (1984) shows that exposure for just 1 second to a sound can result in spectral adaptation, possibly via neural adaptation in the auditory periphery. Summerfield *et al.* (1984) played a sequence of sounds to the listener the first of which consisted of a harmonic complex containing the first 50 harmonics of a tone, with a number of those 50 harmonics reduced in amplitude. The complex was then heard again but with the harmonics which had previously been reduced in level increased to the level of the remaining harmonics. When listening to the second sound, the components that had changed in level were heard to stand out perceptually from the rest of the tone. This enhancement effect appears to demonstrate the auditory system's heightened sensitivity to temporal changes in spectrum and may show a mechanism which helps listeners to perceive differences in magnitude spectra between different sounds, as well as a system that can remove spectral distortion.

The temporal dynamics of the enhancement effect suggest that it is the result of rapid adaptation of neural responses in the peripheral hearing system (Summerfield *et al.*, 1984, 1987). However, the results of onset and recovery measurements are somewhat controversial (Cardozo, 1967; Viemeister, 1980; Wilson, 1970). Therefore, the extent to which the measured temporal dynamics of adaptation are compatible with processing only in the periphery is not clear.

Spectral adaptation may also be caused, at least partly, by central brain processes. Adaptation in the central hearing system appears to be stronger and to have a longer time constant (Holt, 2006; Ulanovsky *et al.*, 2003, 2004; Watkins, 1991). This effect is assumed to occur at the level of the auditory cortex, and it is called *stimulus-specific adaptation*, which is the specific decrease in the response to a frequent (standard) stimulus (Holt, 2006). In addition, such adaptation may also occur at even higher stages in the brain in the cognitive processes. This would mean subconscious calculation of the average spectrum and its inverse filtering during listening. Adaptation may also be inextricably linked to auditory memory, whereby adaptation

occurs by a process of forgetting (McKeown and Wellsted, 2009). These adaptation effects are relatively poorly understood. The time constants in peripheral adaptation and central adaptation are not yet confirmed, and the mechanisms behind adaptation are largely unknown.

11.5.2 Perception of Phase and Time Differences

The Fourier analysis in Equation (3.17a) produces a complex spectrum, which can be presented in polar coordinates using the magnitude and phase

$$\mathcal{F}\{x(t)\} = \mathrm{Re}\,\{X(\omega)\} + \mathrm{j}\,\mathrm{Im}\,\{X(\omega)\} = |X(\omega)|e^{\mathrm{j}\varphi(\omega)}. \tag{11.7}$$

The phase $\varphi(\omega)$ is mathematically problematic, since if it is computed from

$$\varphi(\omega) = \arctan[\mathrm{Im}\,\{X(\omega)\} \,/\, \mathrm{Re}\,\{X(\omega)\}] \tag{11.8}$$

the result is wrapped discontinuously between 0 and 2π. To obtain a continuous function of phase, it has to be *unwrapped*, which is numerically a critical operation and can lead to errors with a magnitude of 2π or its multiples.

The capability of the phase spectrum of sound to transmit information between humans in typical acoustic conditions is now discussed. As an example, the signals in Figure 11.9 have equal magnitude spectra but different phase spectra. In principle, such a prominent change in the temporal structure of the signal could be used to encode some meaning into communicated sounds, such as speech sounds. However, real listening conditions with a relatively distant source contain at least some reflections from nearby surfaces and potentially some reverberation. The impulse response of such a space corresponds to a very complex transfer function,

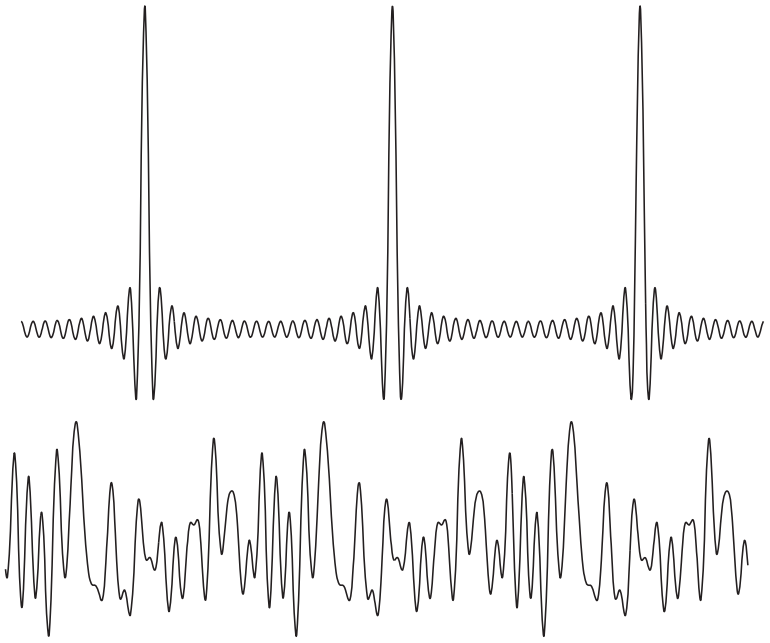

Figure 11.9 Two signals with 40 harmonic partials with identical magnitude spectra. In the upper signal, the phases of the harmonics are synchronized so that the partials have their maxima at the same temporal position. Correspondingly, the phases of the partials are randomly set in the lower signal.

which skews the phase spectrum to almost a random form. Thus, in practice, if the source emitted a signal such as in the upper panel of Figure 11.9, the signal reaching the receiver in a room outside the reverberation radius of the source would look something like the signal in the lower panel. It is thus natural that hearing has developed to be almost immune to phase, although some sensitivity to it exists. In 1843, Georg Ohm proposed that hearing measures the characteristics of a sound as the strengths of its partials and disregards the phase relations between them, which is known as Ohm's acoustic law. This is not completely true, as will be discussed below.

Such insensitivity to phase can often be assumed, although it has been shown that in some cases the phase and a modification of the phase have a prominent effect on perception. In fact, if the signals in Figure 11.9 are listened to with headphones or with loudspeakers closeby (without the effect of the room), a clear difference is perceived. In the upper panel, all partials of the harmonic complex have their maxima aligned at some temporal position, resulting in a prominent, repetitive peak at corresponding positions. In the lower panel, the phase relationships of the harmonics are random, and no such 'peakiness' is seen. The first signal is sharper; that is, it is perceived to have more energy at high frequencies. In addition, the tonal character of such sound has been called 'buzzy' (Moore, 2002), or as having high 'buzzyness' (Laitinen *et al.*, 2013), referring to a low vibrating sound like that of a bee.

Factors leading to a perception of buzzyness

In many cases, when a pair of signals with equal amplitude spectra but different buzzyness is found, a clear difference in the 'peakiness' is seen in the time domain that is, the location and height of peaks occurring in the two signals in the time domain differ. Such a change in peakiness is obvious in the broadband sound signal plotted in Figure 11.9. The peakiness, in this case, is present in the response of the signal at the basilar membrane, as shown in the highest ERB band of the signal, which is plotted in Figure 11.10.

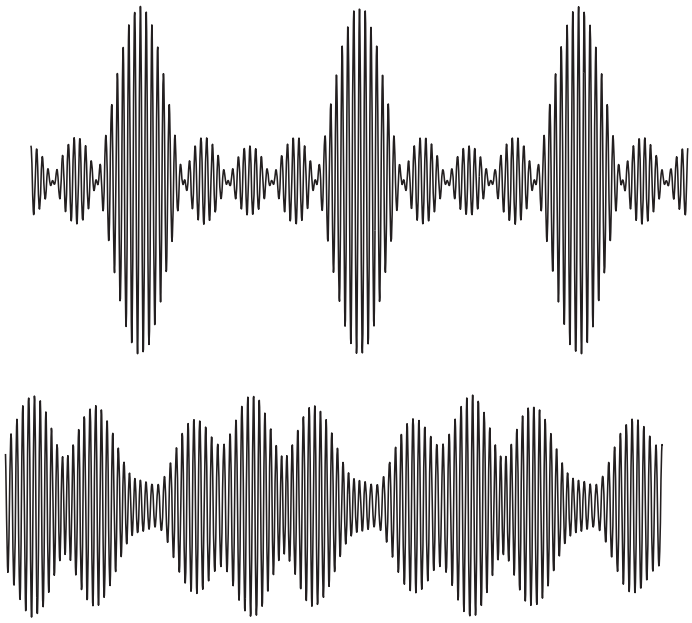

Figure 11.10 The responses of the signals in Figure 11.9 in the ERB band with frequency corresponding to the high end of their spectra.

This difference in peakiness has been studied further by Laitinen *et al.* (2013), where the timbral characteristics produced by various phase alterations are addressed. The study shows that phase randomization of the partials of a sawtooth signal perceptually creates a slightly larger change in timbral characteristics than when the amplitudes of the harmonics are randomized within a standard deviation of 4 dB. In other tests it was found that hearing is sensitive to the peakiness of the signal integrated within bandwidths of about two octaves. This can be simulated using an auditory model which estimates the neural firing rate at different critical bands depending on time. If the firing rate is synchronized in time within neighbouring critical bands, a buzzy character is perceived in the sound.

For an example, signals with equal amplitude spectra but different phase spectra, which have different timbres in headphone listening, can be visualized using the auditory model as shown in Figure 11.11. The signal in the top-left corner is a sawtooth signal with $f_0 = 100$ Hz, and the firing rate is shown below as a spectrogram. The excitation pattern shows vertical stripes, which means that the neurons of the cochlea at all frequencies fire simultaneously, leading to a very buzzy character of sound. The time-domain signal in the top-right corner is otherwise the same as in the top-left corner signal, but the 31st harmonic (at 3100 Hz) is inverted in polarity, making that frequency clearly audible, corresponding to about 6–9 dB amplification of that harmonic with the sawtooth signal (Laitinen *et al.*, 2013). The corresponding firing rate spectrogram shows an added horizontal stripe at that frequency, implying a time-smeared response, which seems to cause added loudness in this case.

The bottom-left corner shows a time-inversed (equivalently, having reversed polarity) signal of the one shown in the top-left corner. The signal creates an auditory event with a differently perceived level of bass, which can be explained by the time scattering of the firing rate there (Laitinen *et al.*, 2013). Correspondingly, the bottom-right corner of the figure shows the same signal, but now with a random phase spectrum. This results in a sound with low buzzyness, which is seen through the lack of vertical alignment of the firing rate peaks in time.

The reader is encouraged to listen to the effects using headphones; however, note that the group delay of the headphone has a dramatic effect on the phenomenon, and the results are not easily reproducible with all headphone models.

Perception of frequency-dependent group delay of audio devices

A technical measure commonly used to describe the temporal-domain characteristics of a signal or a transfer function is the group delay $\tau_g(\omega)$, defined in Equations (3.23a) and (3.23ab) on page 51. It is related to the delay of the components of the envelope (or modulation) of a signal. For example, the phase characteristics of a loudspeaker can be described by group delay. The mean value of the delay has no importance in one-way communication, such as listening to music records. However, if two-way communication is utilized, the group delay should be minimized to avoid latencies. If the group delay changes significantly with frequency, audible degradations of quality may occur.

The JND of group delay has been measured to be about 4–5 ms across the entire audible frequency range, when the group delay changes smoothly with frequency (Patterson, 1987). Such changes in group delay, depending on the frequency, are most audible with impulsive sounds. For example, in perceptual coding of audio, the coding artefacts are often first heard as time-smearing of impulses. The impulsive sounds may be degraded to sounds resembling chirps or noise bursts. The JND is smaller when the group delay changes less smoothly with frequency. In Blauert and Laws (1978), the detection threshold of a change in group delay was

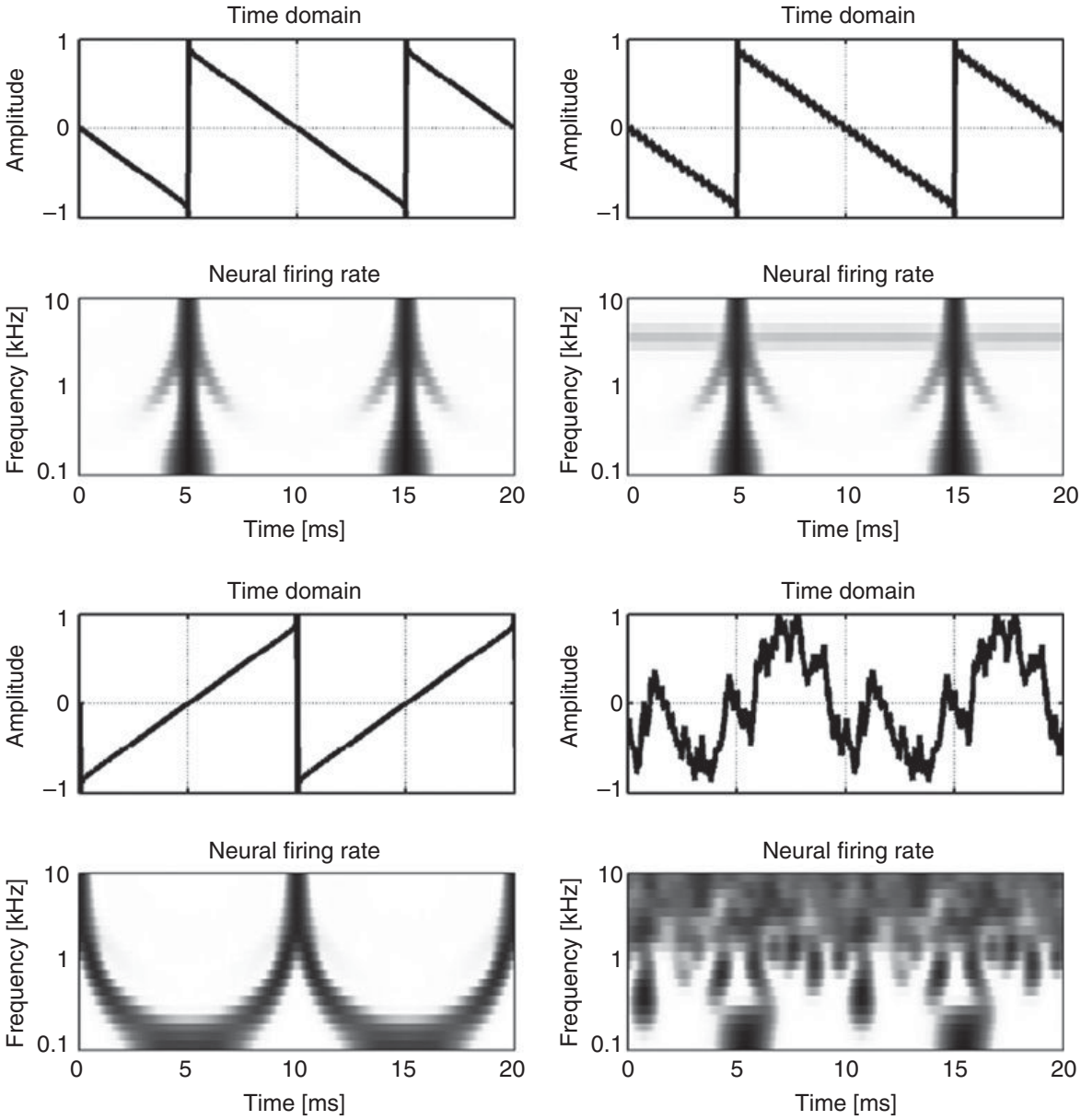

Figure 11.11 Four signals with identical magnitude spectra. The firing-rate panels show a spectrogram type of plot of the instantaneous neural firing rate in the cochlea. See the text for details. Courtesy of Mikko-Ville Laitinen.

as low as 0.4 ms when the frequency dependency of group delay was otherwise smooth, and a local group delay peak about half an octave wide was situated at a certain frequency.

The perceptibility of smoothly frequency-dependent group delays can be described as follows by a set of heuristic rules (Karjalainen, 2008):

- At best, our hearing can detect group delay changes, which are smooth in time, of 1 ms within a critical band. For example, sound signals from the elements of a multi-way loudspeaker at different distances from the listener arrive at different times. In laboratory conditions, delay differences of a little below 1 ms can be detected when directly compared to the non-delayed sound. This means that the distances from the elements of a loudspeaker to the

listener should be similar, not deviating by more than 20–30 cm, which is normally not a problem in loudspeaker design.
- Group delay changes have to be about 3–5 ms with speech and music before they can be perceived. If the changes are of the order of 5–10 ms, they begin to be detectable if the effect of the listening room is minimal. With some phase-insensitive signals, such as noise or diffuse sound, changes in group delay may not be perceived even with longer delays.
- Although the changes in group delay may be perceived as changes in the timbre of sound, the effect on communication, or on speech intelligibility, can be small. For example, even 100-ms group delay changes can be introduced into speech signals before intelligibility starts to vanish.

If the group delay introduced by a communication channel is not smooth with frequency, and has abrupt phase delay differences in adjacent frequencies, the above-mentioned rules no longer apply. One such typical response is reverberation, where, at frequencies over the critical frequency (see Section 2.4.4), the phase response behaves randomly. In practice, most of the phase-related effects are not audible at all in reverberant conditions. When a buzzy signal is filtered with such a random phase response, the buzzyness vanishes, and the firing rate spectrogram resembles the lower-right plot in Figure 11.11. If the phase delay is changed at one frequency in such a case, almost invariably, no audible changes are introduced. In other words, a phase-sensitive signal can be made phase-insensitive by adding the effect of room reverberation to it.

In anechoic listening to buzzy signals, minimal changes in group delays may be audible. This has already been demonstrated with a sawtooth wave, modified by a group delay smooth with frequency, but containing local abrupt change. In the top-left and top-right cases in Figure 11.11, the amplitude spectra are otherwise equal, but the phase of one harmonic has been reversed. This corresponds to applying a group delay of only 0.32 ms at 3100 Hz and none at other frequencies, making that frequency much louder. Such frequency-dependent variations in an otherwise smooth group delay can thus have a large effect during anechoic or headphone listening.

11.6 Psychoacoustic Concepts and Music

Music is a form of art which is composed of sound and silence. It is an art communicated through hearing, and the relations of musical sounds to psychoacoustics are interesting here. The sounds used in music have diverse dimensions, and many of these dimensions are related to psychoacoustics and to the theory of hearing. Some of the phenomena in music can be directly explained by results obtained from psychoacoustics.

This section contains a discussion on two psychoacoustic views of music in two specific perspectives: melody and harmony, and rhythm. Note that in music, terms like consonance, dissonance, and rhythm are very complex and depend on the style of music, and the discussion in this book is therefore limited to only a ‘sensory’ perspective.

11.6.1 Sensory Consonance and Dissonance

A musical note, simply a ‘note’ in the following, is defined as a sound that evokes pitch and has a defined duration. This is a somewhat restrictive definition, since it excludes, for example, drum sounds. However, for simplicity, a note is here defined to be a complex harmonic tone

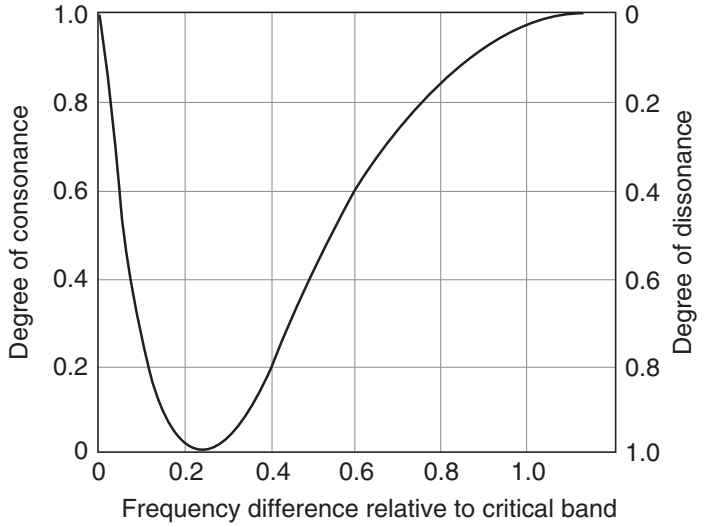

Figure 11.12 The consonance and dissonance levels created by two tones as a function of their frequency difference on a linear scale from zero to the width of the critical band. Adapted from Plomp and Levelt (1965), and reproduced with permission from the Acoustical Society of America.

with a fixed fundamental frequency. Each note with a certain frequency is also often named, for example, C, D, E, F, and so on. The notes may also have a defined duration. Most musical instruments produce a harmonic spectrum when a single note is played on them. Thus, for simplicity, the term 'note' refers to a sound with a harmonic spectrum in this section.

Two notes of different frequencies presented at the same time create a pleasant or an unpleasant sound. If the resulting sound is pleasant, the original notes are said to be in consonance, and correspondingly, for an unpleasant result they are said to be in dissonance. These concepts are central to music, especially Western music, and to the structures to which the discussion in this book is limited.

The level of consonance and dissonance has been measured with listening tests for two sinusoids as a function of the difference in frequency, and the result is presented in Figure 11.12 (Rossing *et al.*, 2001). When the frequencies are the same, consonance has a maximum value of 1.0 and dissonance 0.0. When the frequency difference is about a quarter of the critical band, dissonance has its maximum and consonance is at its minimum. After this, consonance approaches asymptotically the value 1.0 when the frequency difference is increased.

In principle, a single note can produce considerable roughness or dissonance, since, with certain pitches, some of the harmonics at high frequencies may be at a distance of a quarter of a critical band. This may be understood better with the results shown in Figure 11.8, which shows that the separation between partials should be less than 100 Hz to create a noticeable perception of roughness. With harmonic complexes, or notes, this can be achieved only with frequencies below 100 Hz and when at least some of the higher frequency partials have a large amplitude. Typically, instrument sounds have an amplitude that decreases with frequency, and such roughness issues are not encountered. On the other hand, in some synthetic bass sounds such roughness may exist.

The roughness phenomenon is, on the other hand, much more easily obtained when two notes are played at the same time. If the fundamental frequencies are different, the distances of their harmonics on the frequency scale define the roughness. If some of the harmonics are located at distances relative to each other that create roughness, a dissonant sound is perceived,

as shown for three intervals in Figure 11.13. When the frequency ratio is near 3:2, as shown in the uppermost case in the figure, most of the partials of the spectra are either located very near to each other or far enough away from each other to cause no roughness. Only partials with number 10 or higher are at frequencies influencing each other. When the frequency ratio is about 5:4, a larger number of partials are seen to have such a ratio with each other that higher roughness is perceived. With a frequency ratio of $\sqrt{2}$, most partials interfere with each other, and a high level of roughness, and also a high degree of dissonance, is obtained.

Dissonance has been measured for different ratios between the fundamental frequencies of two notes. Figure 11.14 shows the degree of dissonance and consonance when the fundamental frequency of one of the sounds stays at 250 Hz, and the frequency of the other sound changes from 250 Hz to 530 Hz (Plomp and Levelt, 1965). Each sound consists of the fundamental frequency and five harmonic components. Perfect consonance is obtained with frequency ratios 1:1 and 1:2. Consonance has a local maximum with all simple integer ratios, such as 2:3, 3:4, 3:5, and so on, and a minimum in between.

11.6.2 Intervals, Scales, and Tuning in Music

The notes in a *musical scale* differ from each other by fundamental frequencies, which are specified by the tuning system. The notes are at distances on the scale called *intervals*, and they are quite often specified as a ratio between the frequencies. In most scales, an octave corresponds to doubling or halving the frequency. A *melody*, by its simplest definition, is a succession of notes with varying frequencies and temporal lengths. A *chord* is composed of multiple notes presented simultaneously. *Harmony* can then be defined simply as the tonal character of the chords and possibly also their succession in time.

Sensory consonance and dissonance provide the foundation to understanding the structure of melodies and harmonies in music. Different musical styles can be characterized based on their tendency to involve dissonance. For example, choral works from the Renaissance era hardly ever have any dissonant chords; classical music uses dissonance to build tension, which is released in consonant chords; and Jazz music basically avoids consonant chords completely.

The plot in Figure 11.14 helps us understand why the intervals of notes with simple integer ratios are the basis of melodies and harmonies. The *octave* with ratio 2:1; the *fifth* with ratio 3:2; the *fourth*, 4:3; the *sixth*, 5:3; and the *third*, 5:4 are examples of simple intervals. Almost all scales in music utilize such intervals. As a basic rule, the larger the integers in the ratio are, the lower the consonance is.

The diatonic scale divides an octave into seven notes and into repeated octaves. The intervals between adjacent pitches are five whole tones and two semitones. The order of the intervals in the Western major scale has two whole tones followed by one semitone, three whole tones and one semitone, ending at the octave of the first note. The notes in C major are called C, D, E, F, G, A, B, and the one-octave higher C. Each of the notes may also be ‘raised’ (or in some cases ‘augmented’) or ‘lowered’ (or in some cases ‘diminished’) by a semitone. The notation is then, for example, D♭ (D flat) or D♯ (D sharp), respectively. The number of notes is thus 12, assuming that a whole tone is divided into two semitones, resulting in one note in between. This means that C♯ equals D♭; D♯ equals E♭, and so on. The whole tones in the scale do not necessarily correspond to an equal frequency ratio, and the same holds for semitones.

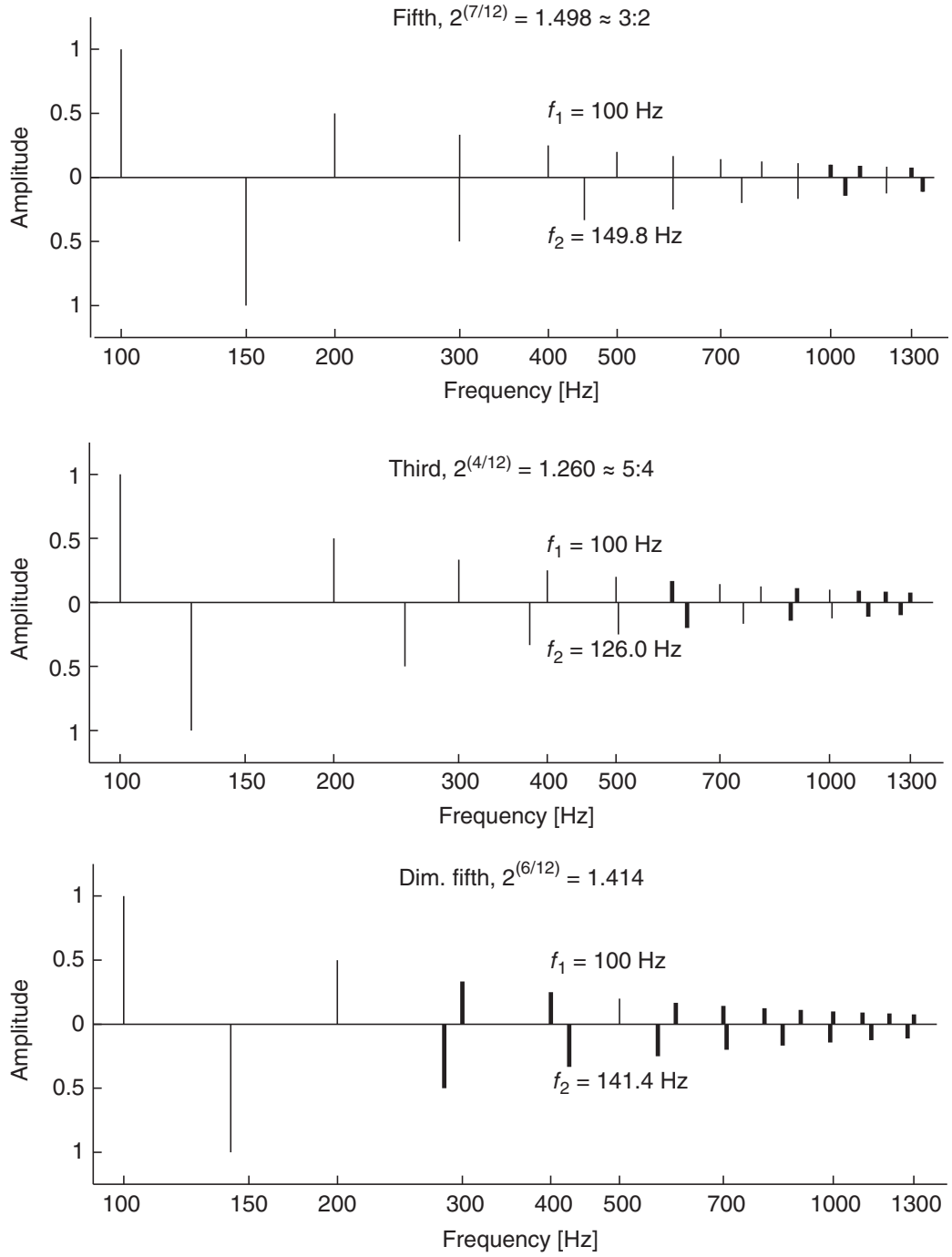

Figure 11.13 The amplitude spectra of three pairs of harmonic tone complexes with the name of the corresponding interval and frequency ratio given in the figures. The abscissa shows the frequency on a logarithmic scale, and the ordinate is the amplitude of each harmonic. The spectrum of a sound with lower pitch is shown with spectral peaks upwards, and correspondingly, a higher pitch is shown with spectral peaks downwards. The bold spectral peaks represent those harmonic components that interact with the components of the other sound, causing roughness and a higher degree of dissonance.

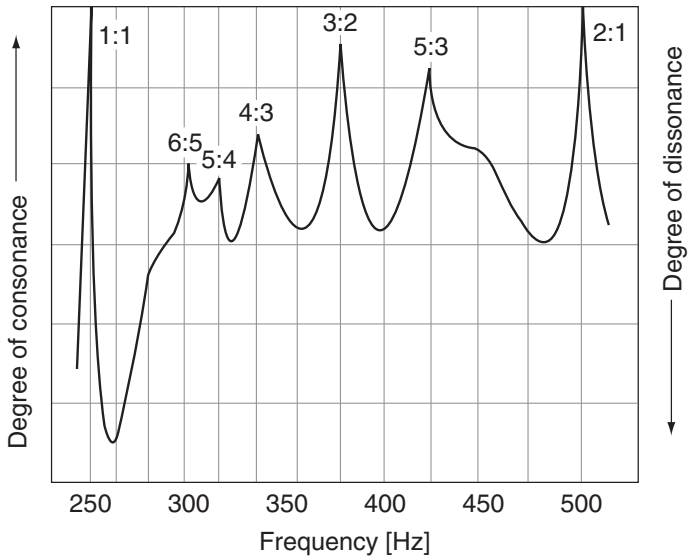

Figure 11.14 The degree of consonance and dissonance created by two harmonic complexes plotted as a function of the frequencies of one of the sounds when the fundamental frequency of the other sound is a constant 250 Hz. Reproduced from Plomp and Levelt (1965) with permission from The Acoustical Society of America.

The frequencies of the notes in the scales then have to be specified, or 'tuned'. The following tuning methods are of theoretical and possibly practical interest:

- *just intonation*;
- *Pythagorean scale*;
- *mean-tone temperament*;
- *equal temperament*.

Just intonation is based purely on *triads*, where the frequencies of the triads C-E-G, G-B-D, and F-A-C have the ratio 4:5:6 and octaves are always tuned to 2:1. Unfortunately, with this method, many intervals are not in tune. For example, the interval D-A should have the ratio 1.5, but in this tuning it is 1.48, which is audibly mistuned.

The Pythagorean tuning system has the greatest number of pure fifths and octaves. Unfortunately, the thirds are then mistuned.

We can thus conclude that no single method exists that tunes the seven notes of the diatonic scale so that all the intervals between the notes have intervals with simple integer ratios. Some methods have been developed to overcome this problem. For instance, the mean-tone temperament alters the Pythagorean tuning so that the fifths are tuned a bit narrower, which overcomes some problems of the system.

Many musical instruments, such as pianos and synthesizers, are nowadays tuned according to the equal temperament system, where a pure octave is divided into 12 semitones having an equal frequency ratio $\sqrt[12]{2} \approx 1.05946$. A whole tone thus corresponds to two semitones $\sqrt[6]{2} \approx 1.225$. The intervals shown in Figure 11.13 use the equal temperament system, where, for example, the interval 'third' should have a frequency ratio of 5:4 (1.25), but the equal temperament system defines the ratio to be 1.26. If the interval ratio was 5:4, the fifth partial of $f_1 = 100$ Hz note would have a frequency of 500 Hz, and the fourth partial of $f_2 = 125$ Hz note

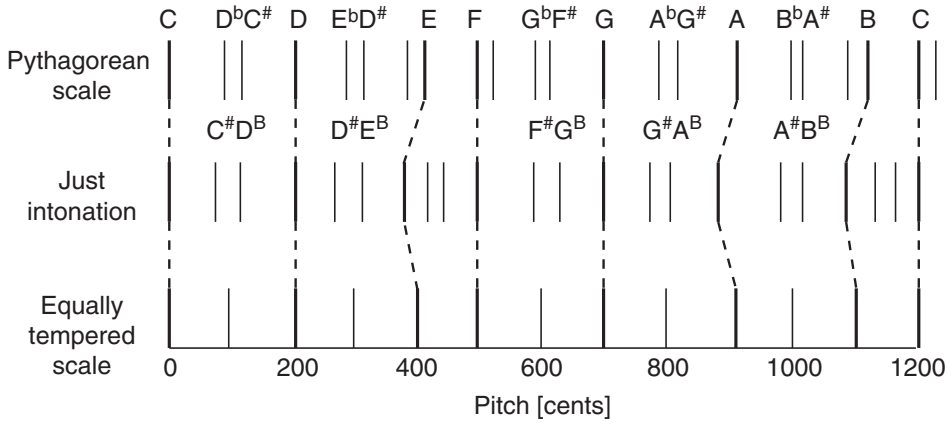

Figure 11.15 The pitches in the Pythagorean scale just intonationand the equally tempered of one octave as a function of logarithmic pitch expressed in cents. Rossing *et al.*, 2001.

would also be 500 Hz, producing no beating or roughness. However, in equal temperament, the fourth partial of $f_2 = 126$ Hz note is located at 504 Hz (as shown in the figure), which interferes with the partial of f_1 note at 500 Hz. However, the difference is so small that roughness is not generated, and only some beating is perceived. A semitone is further divided into 100 *cents*, each cent corresponding to the ratio $\sqrt[1200]{2} \approx 1.0005778$.

The Pythagorean scale, just intonation, and the equally tempered scale are compared in Figure 11.15. The scales clearly exhibit irregular differences. Note that only with equal temperament is the number of individual frequencies 12, since the sharp notes have equal counterparts in flat notes. With the other scales, the sharp and flat notes do not have the same frequencies in any case, i.e., for example, C♯ does not equal D♭.

11.6.3 Rhythm, Tempo, Bar, and Measure

Probably an even more important musical structure than melody and harmony is *rhythm* and its related features. Rhythm is a complex concept which refers to different temporal structures in music. The existence of rhythm is based on natural repetitions in time, such as walking, running, the heartbeat, and breathing. In general, a systematic description of rhythmic structures is harder than one for the structures in melody and harmony. The concepts related to rhythm are also slightly imprecise.

Some other concepts related to rhythm are:

- *Note value*: The relative temporal length of a note. The basic values are the full note 𝅝 ; the half note 𝅗𝅥 ; the quarter note ♩ ; the eighth note ♪ ; and the sixteenth note 𝅘𝅥𝅯. The duration of the notes is halved each time when stepping forwards in the presented list of note values. The lengths of pauses, or silences, between notes are also defined accordingly.
- *Measure* or *bar*: A rhythmic 'placeholder' which indicates a prototype repeated rhythm in music. The number of 'prototype' notes and their duration is shown in the time signature N/M, where N is the number of notes with the value of $1/M$ that can be included in each measure. For example, the time signature $3/4$ means that the temporal length of each measure is such that three quarter notes can be played during the measure as | ♩ ♩ ♩ |. In actual

music, the prototype notes in a bar can be replaced by rests, divided into shorter notes, or joined into longer ones without the temporal length of the measure changing.
- *Tempo*: The speed of presentation. In the notation of written music, ♩ = 80 means that the speed of presentation is 80 quarter notes per minute. The tempo also typically varies during the performance, and the tempo notation is simply a recommendation of the tempo.
- *Beat*: The accenting of specific temporal positions in a bar. In music with a $4/4$ measures, the downbeat denotes that the first note should be emphasized with a milder accent on the third note. The upbeat, in turn, denotes that the second and fourth prototype notes should be emphasized. There are also other meanings for 'beat', but they are beyond the scope of this book.

Performed music quite seldom exactly follows in rhythm such mathematically defined temporal lengths. Musical notation must be taken as a simplified way of communicating and storing musical melodies, rhythms, and harmonies, and musicians must know the style of playing to realize the musical piece in the manner intended by the composer.

In a simple case, a single-voiced melody with a static tempo and a static time signature can easily be described and analysed. Unfortunately, there are plenty of examples of music with more complicated rhythmic patterns, whereby a description cannot be made easily (Fastl and Zwicker, 2007). Sadakata *et al.* (2006) use Bayesian theory to relate the rhythmic patterns of music production to the perception of it.

11.7 Perceptual Organization of Sound

The perception of the auditory environment can either be comprehensive or focus on certain details in it. There appear to be two processes running in parallel during perception. A primitive, bottom-up mechanism automatically orders incoming acoustic stimuli via certain acoustic features. The second, top-down mechanism allows us selectively to attend to whichever features we desire (for instance, a certain pitch, location, time interval, or frequency range). In general, we can focus our attention only on one detail at any one time, although by switching rapidly and by using our short-term auditory memory we can attend to a few targets at the same time. Simultaneously, the top-down processing parses the details and organizes the auditory environment as a whole and tries to focus on some new details for active monitoring.

The hearing mechanism involves certain inborn capabilities to analyse the summed sounds of the auditory environment arriving from multiple sources with or without room reflections and reverberation. The capability to discern the source signals is also partially learned and based on experience of similar situations. Some general principles on how human hearing performs this task will be discussed in the following and are discussed extensively by Bregman (1990).

The organization of sensations and perceptions has been a long-standing topic in experimental psychology. *Sensation* refers to the representation of a real-world object by a sensory organ, whereas *perception* means the higher-level interpretation of the real-world object formed by the brain based on a set of sensations. Perceptions are complete interpretations of objects or events, whose organization does not require exhaustive sensory information about the object being inspected.

Universal laws of pattern formation have been found, particularly in the domain of visual perception. For example, the school of Gestalt psychology has focused on finding the principles of grouping in order to explain the emergence of organized patterns. The most common Gestalt laws of grouping are:

- *Principle of proximity*. When two sensory elements are close to one another both in time and space, they tend to be grouped perceptually.
- *Principle of similarity*. Sensory elements resembling each other tend to be grouped together, while differing elements are thought to belong to another object.
- *Principle of closure*. Complete forms and figures tend to be perceived even if part of the figure is hidden. In the case of a pure tone being interrupted sequentially by bursts of white noise, the human auditory system assumes the pure tone continues uninterrupted during the noise bursts.
- *Principle of continuity*. Continuation of a pattern in time, space, or in some other dimension is a strong assumption, unless contradicting sensory information is presented. For example, the smooth pitch variations and smooth formant changes in speech imply to the listener that the speech originates from the same speaker and is organized into a single stream. If too-rapid changes in pitch or formants are introduced, multiple streams may be created (Plack, 2013).
- *Principle of common motion*. If sensory elements move in the same direction at the same rate, they tend to be grouped as parts of a single stimulus.
- *Principle of belongingness*. Each sensory element can (usually) only belong to a single perceptual object.

Gestalt psychology has many principles that are common to the sensory modalities (vision, auditory, somatosensory, olfactory, and gustatory). Naturally, each of the modalities can have individual implementations of these principles.

The theory and techniques of *pattern recognition* have developed through research into machine vision and machine hearing. In pattern recognition, information captured from the real world (for example, a visual or auditory signal) is pre-processed and transformed into a vector of relevant features. The acquired feature vector is then compared to a set of known feature vectors using a suitable algorithm, with the goal of recognizing (classifying) the unknown pattern. Pattern recognition often requires the formation of quantitative metrics for distance or similarity (compare with the principles of grouping).

11.7.1 Segregation of Sound Sources

We know by experience that an auditory event and the corresponding sound event can usually be related to a specific sound-producing object, the *sound source*. Thus, to a human listener, grouping sound events or their constituents based on the sound sources is more meaningful than basing a grouping on the acoustic properties of the sound events (although they are often interrelated). For example, an approaching car produces engine noise, tyre noise, or a horn signal that are distinctively different sound events. However, when properly aligned, we experience their combination foremost as a unified object, a sound source to which we can attribute a functional meaning thanks to our previous encounters with similar sound sources.

One of the most remarkable capabilities of the human auditory system is *sound segregation*, or *source separation*. A traditional example is the so-called *cocktail party effect*, referring to our ability to focus on a single speaker in a multi-speaker situation, or in strong background noise. Psychophysical tests have found that subjects use many cues in stream segregation in such cases (Bronkhorst, 2000).

Another example is listening to music. An experienced musical listener can distinguish details from complex sound masses with amazing accuracy. However, not everybody gets to

practise their analytical listening skills, because music does not have a communication function similar to speech.

A third example is street noise. We can be somewhat aware of our surroundings by using auditory information only (the blind have to learn this to an even greater degree). Spatial hearing and hearing the space are important components in orienting in the sonic environment.

11.7.2 Sound Streaming and Auditory Scene Analysis

As an example of the perceptual organization of auditory events, we will inspect the formation of a melodic line, which is an important structural factor in Western musical tradition. If we listen to sound events placed properly both in time and frequency, depending on the case, we will perceive one or more *auditory streams* (Bregman, 1990; McAdams and Bregman, 1979). A broader term for the perceptual organization of sound is *auditory scene analysis* (Bregman, 1990).

Figure 11.16 presents a simple example of a repeating sequence of six notes with a relatively wide frequency separation. If we listen to this sequence in a slow tempo, we hear only a single auditory stream (melodic line). When the rate of notes is increased, the higher and lower notes will be separated into two distinct auditory stream, or melodic lines.

Following the principle of belongingness, each note is organized into one of the competing auditory streams (or it can be perceived as a separate stream), but no note is part of more than one streams. Depending on the case, in Figure 11.17 the final note (F) may be organized either into the upper or lower auditory stream. This also changes the perceived rhythmic pattern accordingly.

As the rate of notes being presented increases (and/or the frequency separation becomes wider), the auditory perception is organized into more and more auditory streams, until finally only an ensemble timbre, without any melodic lines, is perceived (Figure 11.18).

It can be stated that the streaming of auditory events, or stream segregation, depends mainly on the rate of notes presented and the width of their frequency separation. Figure 11.19 presents thresholds for coherent auditory streaming, stream segregation, and for the uncertain area

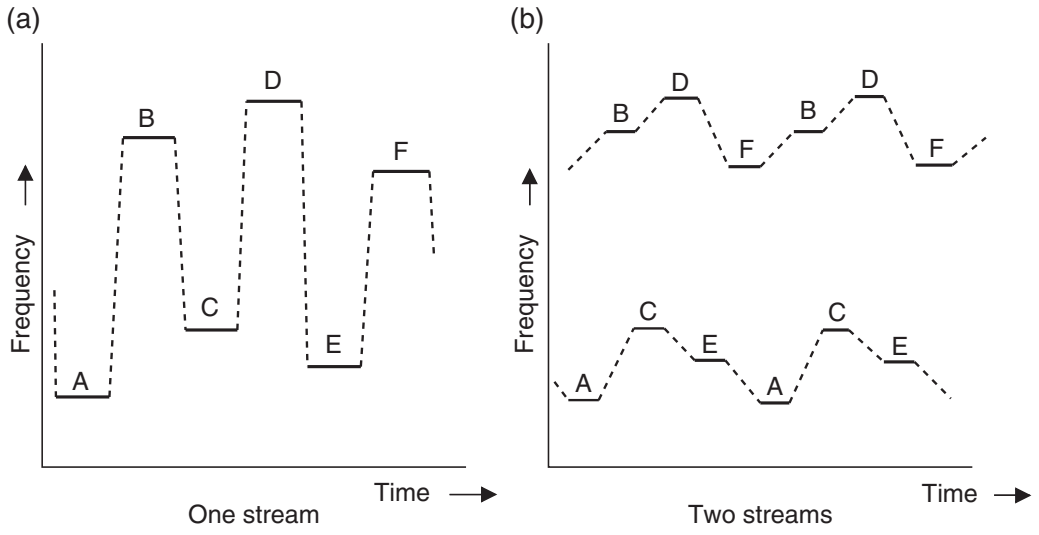

Figure 11.16 The organization of auditory events, depending on the tempo, (a) into one auditory stream for 5 notes per second and (b) into two auditory streams for 10 notes per second.

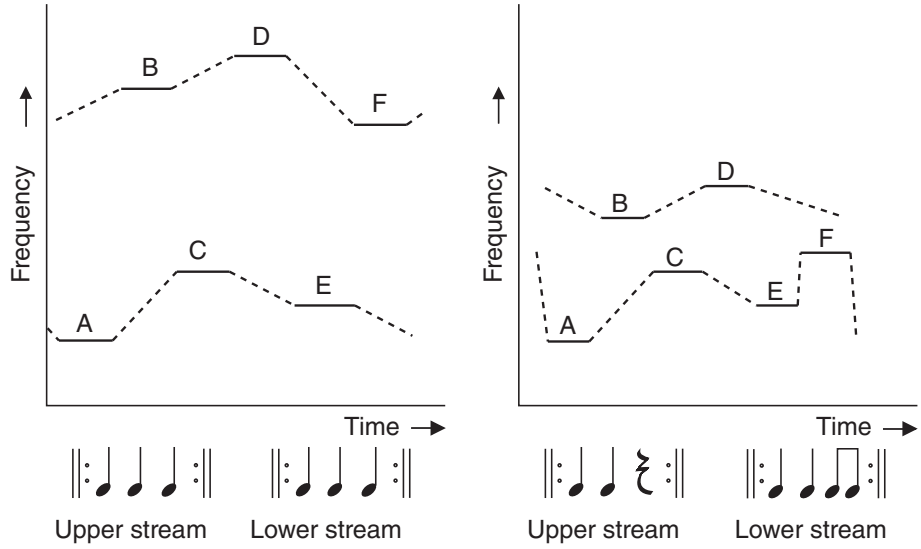

Figure 11.17 A single note (F) may be organized into either of the auditory streams, depending on the case, which also affects the perceived rhythmic pattern.

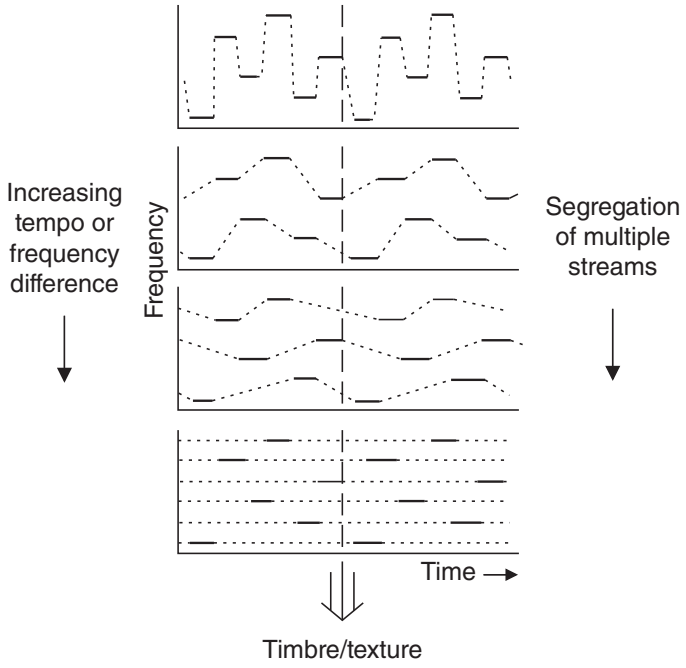

Figure 11.18 As the tempo of the sequence is increased, it begins to segregate into multiple auditory streams, until finally an ensemble timbre, without melodic lines, is perceived.

between them. If the frequency separation of notes is approximately one semitone, auditory streaming occurs regardless of the rate of presentation of notes. Correspondingly, if the temporal interval between notes is more than 150 ms, the probability of the tones being grouped into a single stream is also increased for wide frequency separation between notes. However, this happens in the uncertain area where listeners may perceive either a single coherent auditory stream or segregated streams (the hatched area).

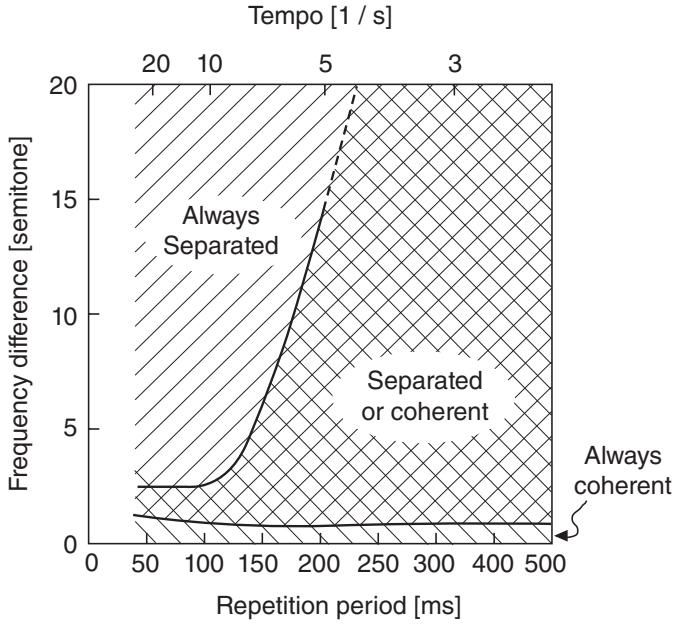

Figure 11.19 Auditory event streaming or segregation as a function of the rate of presentation and frequency separation. Adapted from van Noorden (1975).

Furthermore, factors other than the tempo and frequency separation affect the auditory streaming too. Similarity in the timbre or in some other factor tends to aid stream integration. Additionally, continuation factors such as smooth changes in frequency at the beginning or end of a note may cue that specific note for stream integration.

So far, we have studied only the *sequential streaming* of notes, or, in other words, streaming of auditory events occurring in succession, not simultaneously, in time. Perhaps an even more fundamental concept is the integration of simultaneous auditory components; that is, organizing a set of concurrent components into subsets to represent distinct sources of sound. This is closely related to, for example, perceiving the fundamental frequency of a sound with a harmonic or near-harmonic structure as a coherent whole. Other factors facilitating simultaneous grouping are, for instance, the synchronization of onsets, synchronized amplitude or frequency modulation, and uniform spatial information.

In a sense, auditory scene analysis is a continuation of the fundamentals of psychoacoustics, and it offers future research challenges in understanding and modelling the higher-level properties of the human auditory system. *Computational auditory scene analysis* (CASA) (Wang and Brown, 2006) is a promising approach to solving the puzzle.

Summary

This chapter complemented the discussion on psychoacoustic quantities by introducing sharpness, fluctuation strength, impulsiveness, tonality, and roughness, which can all be seen to be subcategories of timbre. The quantities are interesting when trying to understand what kind of signals draw the attention of the listener. Many of the quantities have a connection to human sensitivity to modulations in sound. We seem to be tuned to pay attention to amplitude and frequency modulations in the range from about 1 Hz to 16 Hz. In turn, modulations at 30–100 Hz result in the perception of rough and unpleasant sound.

Perception of the magnitude response and phase spectrum variations has also been described; this clearly shows that humans are sensitive, in some cases, to the phase spectrum of a signal. Time-domain peaks seem to cause the perception of buzzyness, but if the sound source is in a room, the response smooths out such peaks in many cases.

A clear connection between psychoacoustics and music has been shown. Roughness is closely related to consonance and dissonance perceived between two or more musical notes; the more the partials of the notes generate roughness, the higher the perceived dissonance is. This leads directly to basic melodic and harmonic structures in music.

The perceptual organization of complex auditory scenes into streams was also described in this chapter. The laws of pattern formation describing how the sensory information is grouped into auditory objects were discussed, and some illustrative cases in music shown.

Further Reading

The fundamentals of the psychoacoustic quantities discussed in this chapter are outlined by Fastl and Zwicker (2007). A good introduction to music acoustics is provided by Rossing *et al.* (2001), and the primary source for auditory scene analysis is (Bregman, 1990).

References

Bienvenue, G.R. and Nobile, M.A. (1991) Prominence ratio for noise spectra with discrete tones: A procedure based on Zwicker's critical band research *Proc. of Inter-Noise*, **1**, 53–55.

Blauert, J. and Laws, P. (1978) Group delay distortions in electroacoustical systems. *J. Acoust. Soc. Am.*, **63**, 1478–1483.

Bregman, A. (1990) *Auditory Scene Analysis*. MIT Press.

Bronkhorst, A.W. (2000) The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acta Acustica United with Acustica*, **86**(1), 117–128.

Cardozo, B.L. (1967) Ohm's law and masking. *IPO Annual Progress Report*. Institute of Perception Research, Eindhoven The Netherlands, **2**, 59–64.

Daniel, P. and Weber, R. (1997) Psychoacoustical roughness: Implementation of an optimized model. *Acta Acustica United with Acustica*, **83**, 113–123.

Fastl, H. and Zwicker, E. (2007) *Psychoacoustics – Facts and Models*. Springer.

Holt, L.L. (2006) The mean matters: Effects of statistically defined nonspeech spectral distributions on speech categorization. *J. Acoust. Soc. Am.*, **120**(5), 2801–2817.

ISO-7779 (2010) Acoustics – measurement of airborne noise emitted by information technology and telecommunications equipment. International Organization for Standardization.

Johnston, J.D. (1985) Transform coding of audio signals using perceptual noise criteria. *IEEE J. Selected Areas in Commun.* **6**(2), 314–323.

Karjalainen, M. (2008) *Kommunikaatioakustiikka* (in Finnish), *Communication Acoustics*. Teknillinen korkeakoulu.

Laitinen, M-V., Disch, S., and Pulkki, V. (2013) Sensitivity of human hearing to changes in phase spectrum. *J. Audio Eng. Soc.*, **61**(11), 860–877.

McAdams, S. and Bregman, A. (1979) Hearing musical streams. *Computer Music J.*, **3**(4), 26–60.

McKeown, D. and Wellsted, D. (2009) Auditory memory for timbre. *J. Experimental Psych.: Hum. Percep. Perform.* **35**(3), 855.

Moore, B.C. (2002) Interference effects and phase sensitivity in hearing. *Philosoph. Trans. Royal Soc. London. Series A: Mathematical, Physical and Engineering Sciences*, **360**(1794), 833–858.

Olive, S.E., Schuck, P.L., Sally, S.L., and Bonneville, M. (1995) The variability of loudspeaker sound quality among four domestic-sized rooms *Audio Engineering Society Convention 99*.

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

Pedersen, T.H. (2001) Objective method for measuring the prominence of impulsive sounds and for adjustment of LAeq. *Proc. Int. Congr. and Exhibition on Noise Control*.

Pike, C., Brookes, T., and Mason, R. (2013) Auditory adaptation to loudspeakers and listening room acoustics *Audio Engineering Society Convention 135* AES.

Plack, C.J. (2013) *The Sense of Hearing*. Psychology Press.

Plomp, R. and Levelt, W.J. (1965) Tonal consonance and critical bandwidth. *J. Acoust. Soc. Am.*, **38**, 548–560.

Roederer, J.G. (1975) *The Physics and Psychophysics of Music: An Introduction*. Springer.

Rossing, T.D., Moore, F.R., and Wheeler, P.A. (2001) *The Science of Sound*, 3rd edn. Addison-Wesley.

Sadakata, M., Desain, P., and Honing, H. (2006) The hearing-aid speech quality index (hasqi) version 2. *Music Perception: Interdisc. J.*, **23**(3), 269–288.

Summerfield, Q., Haggard, M., Foster, J., and Gray, S. (1984) Perceiving vowels from uniform spectra: Phonetic exploration of an auditory aftereffect. *Percep. Psychophys.* **35**(3), 203–213.

Summerfield, Q., Sidwell, A., and Nelson, T. (1987) Auditory enhancement of changes in spectral amplitude. *J. Acoust. Soc. Am.*, **81**(3), 700–708.

Terhardt, E., Stoll, G., and Seewann, M. (1982) Algorithm for extraction of pitch and pitch salience for complex tonal signals. *J. Acoust. Soc. Am.*, **71**(3), 679–688.

Terhardt, E., Stoll, G., and Seewann, M. (1996) Pitch of complex signals according to virtual pitch theory: Examples and predictions. *J. Acoust. Soc. Am.*, **55**, 671–678.

Toole, F.E. (2006) Loudspeakers and rooms for sound reproduction - A scientific review. *J. Audio Eng. Soc.*, **54**(6), 451–476.

Ulanovsky, N., Las, L., and Nelken, I. (2003) Processing of low-probability sounds by cortical neurons. *Nature Neurosci.*, **6**(4), 391–398.

Ulanovsky, N., Las, L., Farkas, D., and Nelken, I. (2004) Multiple time scales of adaptation in auditory cortex neurons. *J. Neurosc.*, **24**(46), 10440–10453.

van Noorden, L. (1975) *Temporal Coherence In the Perception of Tone Sequences*. Institute for Perceptual Research.

Viemeister, N.F. (1980) Adaption of masking. In van den Brink, G. and Bilsen, F.A. (eds) *Psychophysical, physiological and Behavioural Studies in Hearing*. Delft University Press, pp. 190–199.

Wang, D. and Brown, G.J. (2006) *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley-IEEE Press.

Watkins, A.J. (1991) Central, auditory mechanisms of perceptual compensation for spectral-envelope distortion. *J. Acoust. Soc. Am.*, **90**(6), 2942–2955.

Wilson, J.P. (1970) *Frequency Analysis and Periodicity Detection in Hearing*. Sijthoff, Leiden, The Netherlands, 303–318.