12

Spatial Hearing

During our evolution, the ability to locate the source of a sound has been critical for our survival. Since the appearance of mammals as primarily nocturnal animals more than 200 million years ago, they have relied heavily on their sound localization abilities for locating friends, foes, and food. Locating the source of a sound remains an important sensory ability for prey and predator alike. Even for humans in modern times, spatial hearing is extremely useful and important for orientation in one's environment. Sound sources are often localized with relatively good accuracy by our hearing, even in cases when the sources are not visible.

The mechanisms of how localization is performed are, however, not generally understood by the layperson. Many of them have been uncovered largely by science. There is actually a plethora of complex, robust, and accurate mechanisms for spatial hearing that are based on signal analysis of either binaural or monaural inputs. For example, the JND in the detection of delays between binaural signals is of the order of 20 μs, which is amazingly accurate when one remembers that it is obtained using neurons whose latency times and output spike lengths are of the order of 1 ms.

Spatial hearing develops substantially through learning and adaptation to gain more accuracy and better performance in complex environments. The fundamental role of learning is easy to understand because spatial hearing is dependent on individual factors, such as the size and form of the head and geometries of the pinnae. The auditory system learns to analyse sound environments by utilizing the properties of direct sound, reflections from surfaces and objects, and reverberant sound arriving at the two ear canals of the subject.

12.1 Concepts and Definitions for Spatial Hearing

12.1.1 Basic Concepts

We begin the discussion by defining some basic concepts related to spatial hearing. The term *localization* is the process by which the location of an auditory event in the auditory space is associated with the attributes of a sound event in an acoustic environment. In general, the human auditory system represents the external sound environment by an internal auditory

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

image or scene (Bregman, 1990). The localization of a sound source can be described by the perception of its direction, distance, and spatial extent. Basically, all auditory events are 'localized', implying that a more or less precisely defined position in the surrounding world or inside the listener's head is associated with them. Reflected and reverberated sound can also be localized, which may lead to perception of the attributes of the geometry of the room and the acoustic properties of the surfaces in it.

The two following concepts are useful in spatial hearing:

- *Monaural hearing* refers to listening in conditions where there is no interaural difference information, or where such differences are ignored. The simplest case of monaural hearing is when the same signal enters both ears (see the definition of diotic listening below).
- *Binaural hearing* means listening to sounds where information due to interaural differences exists and is taken into account.

Spatial hearing is often studied using arrangements enabling the signals entering the two ears to be controlled separately. This is easily possible by using headphones. For such experimental studies, it is useful to define the following concepts:

- *Monotic listening* is a situation where a signal is fed to one ear only.
- *Diotic listening* is the case where a signal is fed equally to both ears.
- *Dichotic listening* is the arrangement where the two ears receive different sounds (which can originate from the same sound source but are processed differently for each ear.)

Note that these concepts do not characterize explicitly the sound-generation techniques used – monophonic, stereophonic, binaural, or multi-channel techniques.

Headphones offer the opportunity for dichotic listening, where the stimulus to each ear can be controlled separately. This is why there have been many studies of dichotic listening with the binaural cues controlled independently. Figure 12.1 illustrates the set-ups for ITD and ILD adjustment using a) controllable delays and b) level attenuators.

Unless special techniques are used, headphone listening typically results in *inside-the-head localization*. In *lateralization*, inside-the-head, localized auditory events are controlled in the left–right direction by modifying interaural differences. Lateralization experiments with

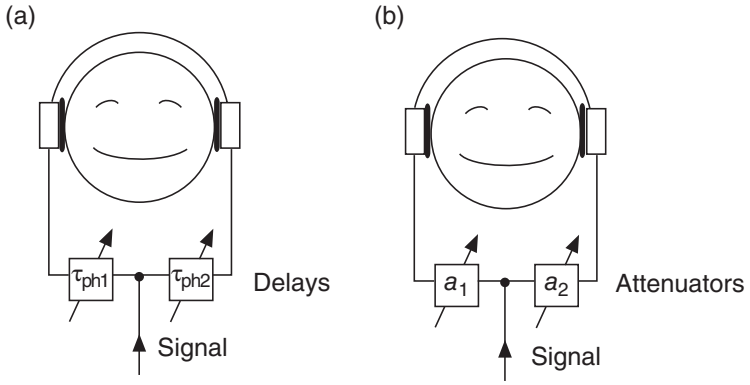

Figure 12.1 Headphone listening set-ups for auditory event lateralization experiments: (a) control of time-delay difference and (b) control of signal-level difference.

headphones are interesting from a theoretical point of view due to the possibility of processing the ear signals independently. On the other hand, such lateralized reproduction often sounds very artificial and does not necessarily reveal much about spatial perception in the real world.

12.1.2 Coordinate Systems for Spatial Hearing

In general, spatial hearing is a three-dimensional phenomenon and therefore requires a three-dimensional coordinate system to describe the acoustic environment. The same spatial information can be mathematically expressed in several different coordinate systems. In this context we define three coordinate systems, each of which has its specific strengths:

1. *Rectangular coordinates* or *Cartesian coordinates* $\{x, y, z\}$. This is a natural way to describe three-dimensional information when observing a sound space and a listener from an external point of view. The listening subject can be positioned at the origin, looking, for example, along the x-axis. From the perceptual point of view of the listener, the rectangular coordinate system is not the most intuitive one.
2. *Spherical coordinates* $\{\varphi, \delta, r\}$ from the most natural 'head-related' or 'listener-centric' coordinate system for subjects when orienting themselves in an acoustic environment. Figure 12.2 illustrates this case using the azimuth angle φ of a direction, $-180° \leq \varphi < +180°$; the elevation angle δ of a direction, $-90° \leq \delta < +90°$; and the distance r of an object.
3. *Cone of confusion coordinates* $\{\varphi_{cc}, \delta_{cc}, r\}$. The geometry of the head and external ear implies a concept called the cone of confusion, discussed below, which is that set of directions where direction discrimination is relatively difficult due to symmetry. This set of directions forms the surface of a cone, hence the name. The cone of confusion coordinate system, characterized in Figure 12.3, is suitable for describing sound source positions from this point of view. The angle φ_{cc} covers the range $-90° \leq \varphi_{cc} < +90°$ and δ_{cc} the range $-180° \leq \delta_{cc} < +180°$.

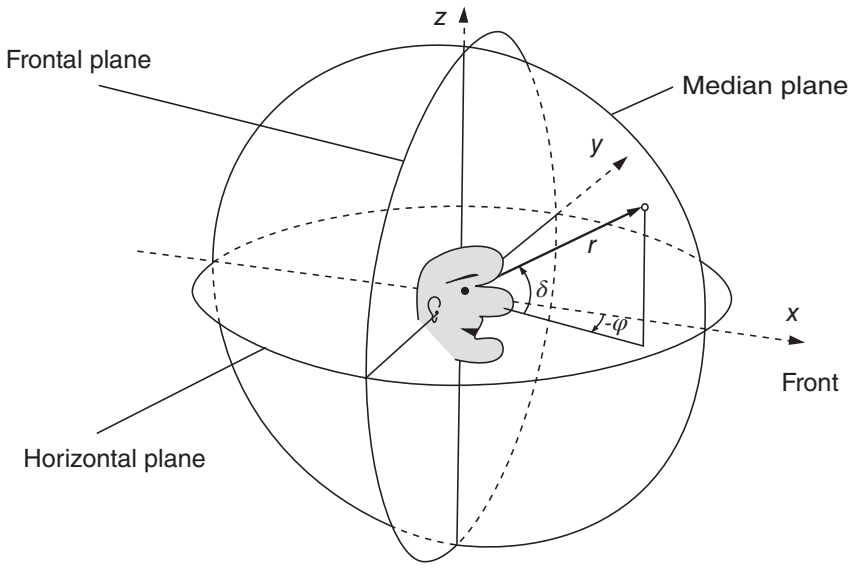

Figure 12.2 The spherical coordinate system, which is natural from the perceptual point of view of a listener. The angle φ is the azimuth and δ the elevation. The distance to the source is r. The three main planes characterized by the circles are the median plane, the horizontal plane, and the frontal plane.

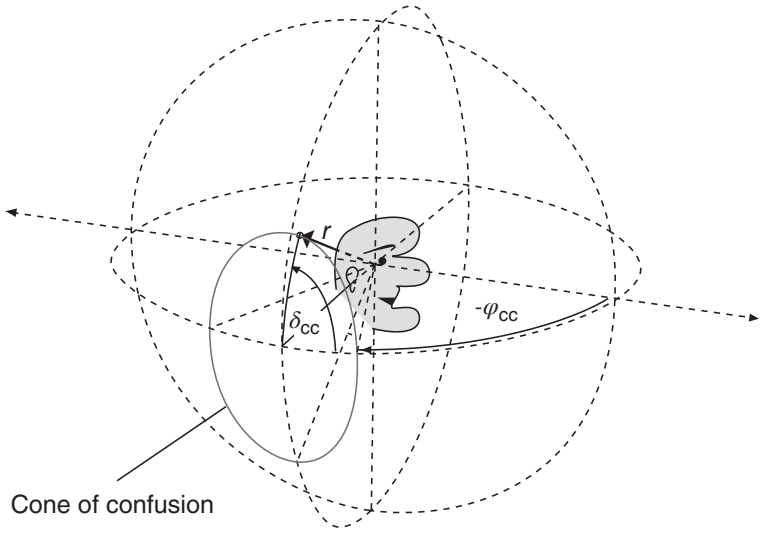

Figure 12.3 The cone of confusion coordinate system, which is motivated by the approximate acoustic symmetry of the head. The angle φ_{cc} is the azimuth angle to the cone of confusion in the horizontal plane and δ_{cc} is the angle on the cone of confusion. The distance to the source is r.

Three planes of symmetry defined below are indicated in Figure 12.2:

- *Frontal plane* $x \equiv 0$ divides the space in the front–back direction.
- *Median plane* (or *median sagittal plane*) $y \equiv 0$ where the distance to the ears is equal.
- *Horizontal plane* where $z \equiv 0$ and $\delta \equiv 0$.

Sometimes, reporting the results from binaural hearing studies requires that the side relative to the ear is specified. The two following terms are often used for this purpose:

- *Ipsilateral* refers to the side of the head that is being discussed.
- *Contralateral* refers to the side of the head opposite to that being discussed.

12.2 Head-Related Acoustics

As discussed in Section 7.1, the transfer function of sound from the source to the ear canal entrance depends on the position of the sound source and the acoustic properties of the listener's head. The azimuth angle φ of the sound source direction, if different from zero, causes interaural time and level differences in the signals entering the ear canals. The torso, particularly the shoulders, also adds reflections. The asymmetry of the head in the front–back and the up–down directions contributes to the difference in signals from different directions.

The *pinna* (Figure 12.4a) makes important contributions to the front–back and up–down differences at frequencies above 4 kHz (Blauert, 1996). The cavities resonate at specific frequencies and affect signals in the ear canal depending on the angle of incidence of sound from a sound source. In particular, the perception of elevation of a source in the median plane, where the ear canal signals are practically equivalent, is aided by the direction-dependent filtering of sound by the pinna.

The combined effect of the torso, the head, and the external ear acoustics is compactly represented in *head-related transfer functions* (HRTFs) or the corresponding *head-related impulse*

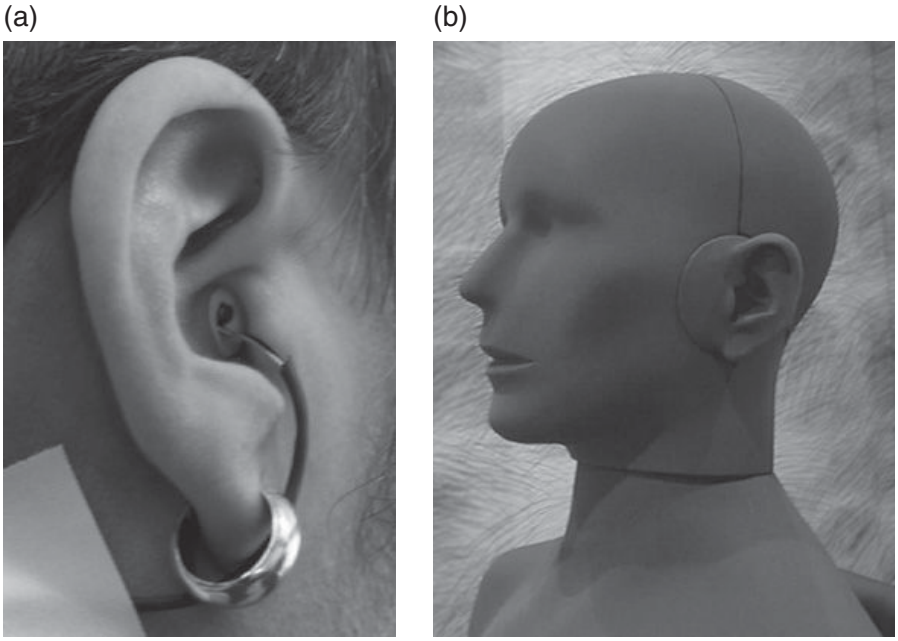

Figure 12.4 (a) The pinna of a subject. The ear canal is blocked by a miniature microphone to measure the head-related transfer function. (b) A dummy head (Cortex MK2) used for binaural measurements and recordings.

responses (HRIRs). Note that the term 'head-related transfer function' is often used generically even when referring to an impulse response.

The head-related transfer function H_{HRTF} is defined by the response to a sound source at a specified position in the ear canal H_{ec} normalized to the response H_{ff} in the middle position of the head when the head is absent:

$$H_{\mathrm{HRTF}}(\omega) = H_{\mathrm{ec}}(\omega) \,/\, H_{\mathrm{ff}}(\omega) \tag{12.1}$$

HRTFs or HRIRs are measured in free-field conditions with the sound source (a loudspeaker) placed in desired positions, for example, 2 metres from a subject at each azimuth and elevation angle of interest (Møller, 1992). An anechoic room can be used for the measurement, but the reflections from walls can also be eliminated by windowing the response to before the arrival of the first reflection. Small microphones are used to capture the response close to the entrance or inside the ear canal. In the former case, the ear canal is typically blocked by the microphone set-up, as illustrated in Figure 12.4a, or it can be kept open. In the ear canal, the measurement position can be any specified point, including the immediate vicinity of the tympanic membrane, in which case the microphone probe must be inserted carefully to avoid damage to the membrane.

In some cases it is practical to use *dummy heads* in binaural response measurements and recordings (Møller, 1992). The dummy heads are also known as *head and torso simulators*, *artificial heads*, or *binaural microphones*. They are designed to approximate the head-related acoustics of a typical human subject. Figure 12.4b shows a dummy head with a removable pinna unit, where condenser microphones are used in place of the ear drums.

A set of HRTFs measured from both ears for three directions is shown in Figure 12.5, and the corresponding HRIRs are shown in Figure 12.6. It can be seen that when the source is in

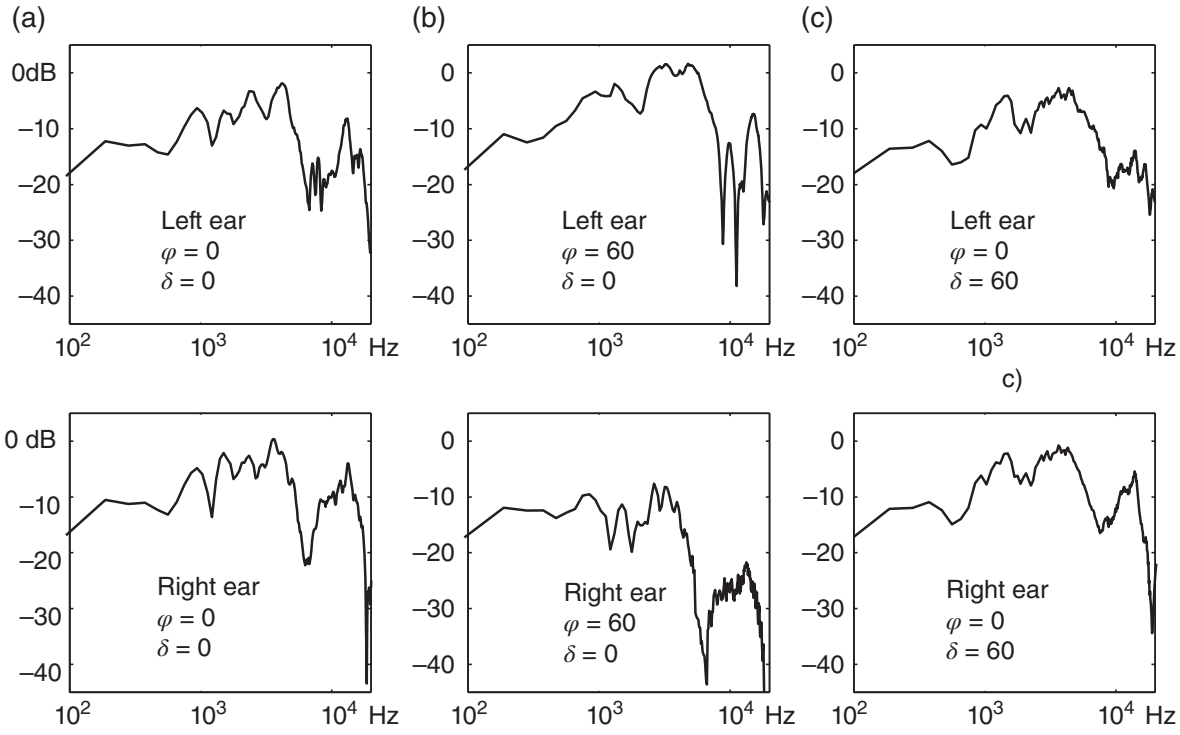

Figure 12.5 A set of HRTF magnitude responses measured from a subject at the blocked ear canal entrance for different directions of sound incidence: (a) $\varphi = \delta = 0°$ (b) $\varphi = 60°, \delta = 0°$, (c) $\varphi = 0°, \delta = 60°$. *x*-axis: frequency. *y*-axis: magnitude in dB.

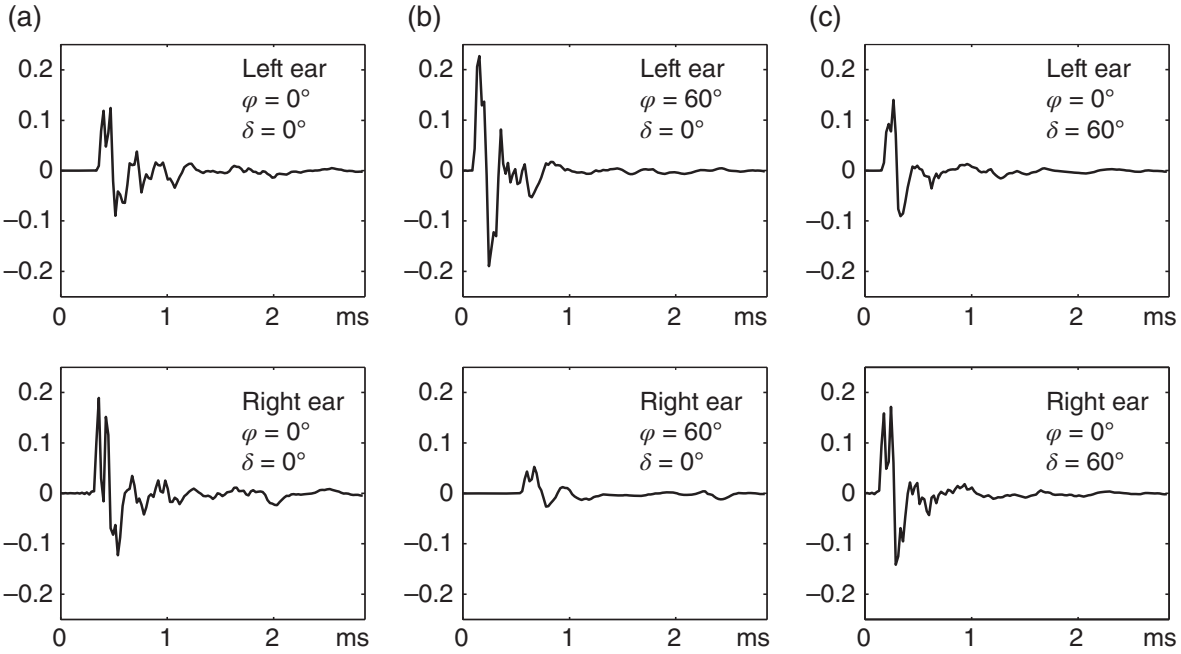

Figure 12.6 A set of HRIRs measured from a subject at the blocked ear canal entrance for different directions of sound incidence (a) $\varphi = \delta = 0°$ (b) $\varphi = 60°, \delta = 0°$, (c) $\varphi = 0°, \delta = 60°$. *x*-axis: time. *y*-axis: sound pressure (relative amplitude).

front, the HRTFs and HRIRs are similar in both ears, which is due to left–right symmetry in humans. Only at high frequencies are there some irregular differences in the responses. The HRTFs are relatively flat up to about 1 kHz, after which there is a broad hump up to about 5 kHz and a dip around 8 kHz, as indicated in Figure 7.3 on page 113.

When the source is moved to the azimuth direction, $\varphi = 60°$, the ipsilateral HRTF shows a boost at high frequencies when compared to the contralateral HRTF. Correspondingly, the direct sound in the ipsilateral HRIR arrives earlier and has a higher amplitude, while the sound in the contralateral HRIR is delayed due to the longer distance it must travel and has lower amplitude due to head shadowing. When the source is in the median plane with elevation $\delta = 60°$, the difference between the contralateral and ipsilateral HRIRs and HRTFs is negligible due to symmetry. An interesting question is how localization is performed, since the frontal responses ($\varphi = 0°, \delta = 0°$) appear to be similar to the elevated case ($\varphi = 0°, \delta = 60°$). This is analysed further below with a larger set of measured responses.

To demonstrate better their dependency on direction, a large number of HRTFs and HRIRs are plotted as 2D surfaces for different azimuths and elevations. Figure 12.7a is a plot of HRIRs measured with a source at 72 directions in the horizontal plane from the left ear of a subject. The time of arrival of the direct sound varies with direction, which is because the centre position of the head of the subject is at the centre of a circle along which the source moves, leaving the ear out of that position by about 8 cm. Thus, the distance between the source and the ear canal changes when the source moves. The source is closest to the ear when in the direction of about 90°, where the earliest start of the response is seen. When the source is on the contralateral side, the response has peaks with lower amplitudes which appear later in time than on the ipsilateral side. An interesting phenomenon is seen at about 1.3 ms for the source direction −100°. The crests of the wavefronts cross, forming a 'bright' spot at the crossing in the 2D visualization of the response. This is because the sound arrives at the ear not only along the shortest route, but also by diffracting around the head along all other paths. This causes multiple peaks in the response, and when the most prominent contributions meet at the ear canal, a higher response occurs.

The corresponding horizontal-plane HRTFs are plotted in Figure 12.7b as a function of the azimuth angle φ. When the source is on the contralateral side, high frequencies are clearly attenuated, which is shown as the black area in the plot. When the source is on the contralateral side in directions from −100° to −110°, the 'bright spot' phenomenon discussed above causes

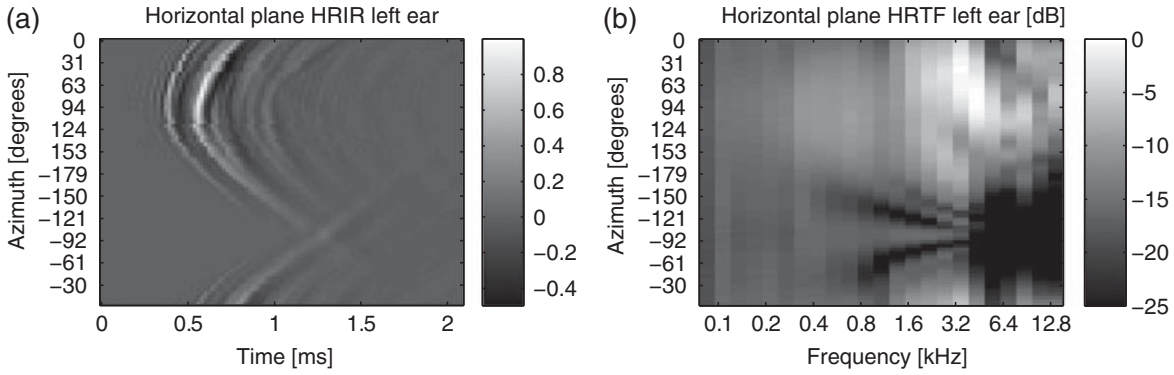

Figure 12.7 (a) The head-related impulse responses measured from the left ear of the subject. (b) The head-related transfer function from the same measurement. The HRTF data originate from Gómez Bolaños and Pulkki (2012).

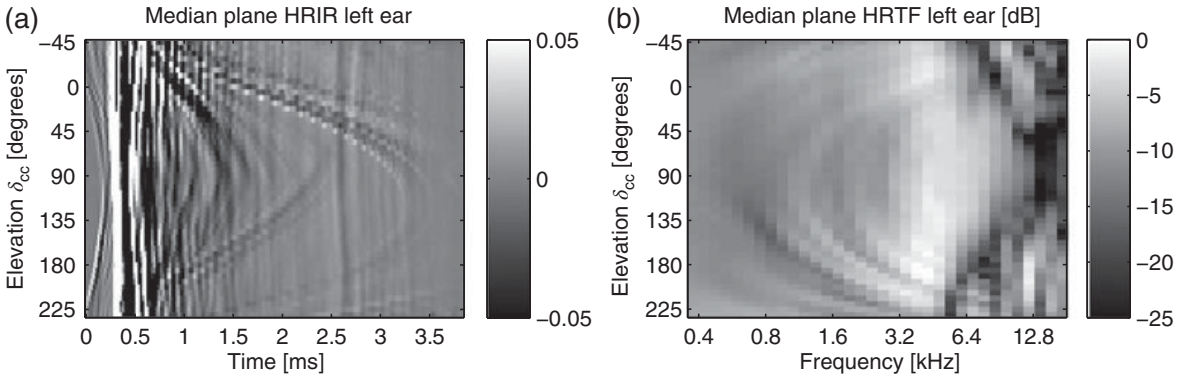

Figure 12.8 (a) HRIRs and (b) HRTFs measured from the left ear of the subject for 72 directions in the median plane. The HRTF data originate from Algazi *et al.* (2001). To emphasize the structure of the HRIR responses after 1 ms, the colour code saturates at ±0.05, although the maximum value of the HRIRs has been scaled to unity.

an amplified response compared to adjacent source directions. The effect is most pronounced at frequencies 1–4 kHz.

The change of HRIR in the median plane is shown in Figure 12.8a as a function of the cone of confusion elevation δ_{cc}. The peaks caused by the sound arriving along the direct path are located at temporal positions 0.3–0.5 ms, which means that the ear canal is relatively well at the centre of the rotation of the source. Perhaps the most interesting details in the figure are the faint arcs or half-arcs which coincide at 0.5 ms with source directions −45° and 225°. The arcs or half-arcs seem to have their apexes at temporal positions of about 1.5 ms, 2.5 ms, and 3.5 ms. The first arc, with apex at 1.5 ms, is caused by the reflection from the shoulders. When the source is above the subject, the sound reaching the shoulders is reflected back to the ears, and thus it travels a distance about 30 cm longer than the direct sound, which causes the 1-ms delay seen in the figure. When the source is lower in the median plane, the extra travel time is shorter, and the shoulder contribution arrives earlier. The reason for the appearance of the other arcs in the figure is not so evident. Very probably they are due to similar systematically changing reflections, for instance from the measurement devices, the chair, or from the feet of the sitting subject. Although the reflections seem quite faint in the figure, they have a noticeable effect on the magnitude response, as will be shown below.

The effect of elevation δ_{cc} in the median plane on the HRTF is depicted similarly in Figure 12.8b. The response depends on the angles, especially above 1 kHz. For example, the frequency of the dip around 8 kHz varies as a function of the elevation angle in the median plane. The arcs at frequencies below 4 kHz with the apex pointing left are indeed caused by reflections from the subject, which manifest themselves as the faint arcs in the corresponding HRIR response. At frequencies above 4 kHz, the response contains a lot of irregularities, which are due to the complex spectral filtering by the pinna and will be discussed below in more detail.

12.3 Localization Cues

The ability of humans to localize sound sources is surprisingly good, especially when you consider that it is based on real-time analysis of two signals entering the ear canals.

Localization is determined or guided by *localization cues* (Grantham, 1995). These are properties of auditory stimuli that are relevant to spatial perception. The combined effect of all cues, weighted by their dominance and relevance, then forms the spatial attributes of subjective auditory events. This section describes the cues that localize the direction of the sound source, the directional cues.

Any physical aspect of the acoustic waveforms reaching a listener's ears that is altered by changes in the position of the sound source may be considered a potential cue in localization. The sound signal undergoes linear distortion, or changes in spectral or temporal content, when propagating from a source to the listener's ears. This distortion forms the basis of most acoustic cues. The cues are commonly grouped as binaural (or interaural) and monaural cues (Blauert, 1996).

Binaural cues are derived from the differences in the signals between the two ears. The two binaural cues, which are also the main acoustic cues in general, are the *interaural time difference* (ITD) and the *interaural level difference* (ILD). The ILD is also often called the *interaural intensity difference* (IID). Particularly in the horizontal plane, these cues have been found to be dominant in localization.

12.3.1 Interaural Time Difference

The *Interaural time difference* (ITD) occurs due to the finite velocity of sound and differing distances from the source to the ears. Figure 12.9 illustrates the principle of ITD formation for a spherical head model and the dependence of the ITD value on the azimuth angle φ in the horizontal plane, computed from a simple, yet relatively accurate approximation rule:

$$\tau = \frac{D}{2c}(\varphi + \sin\,\varphi), \tag{12.2}$$

where φ is the azimuth direction of the source in radians, D is the diameter of the head, and c is the velocity of sound. This rule is easily derived from the spherical geometry of the head.

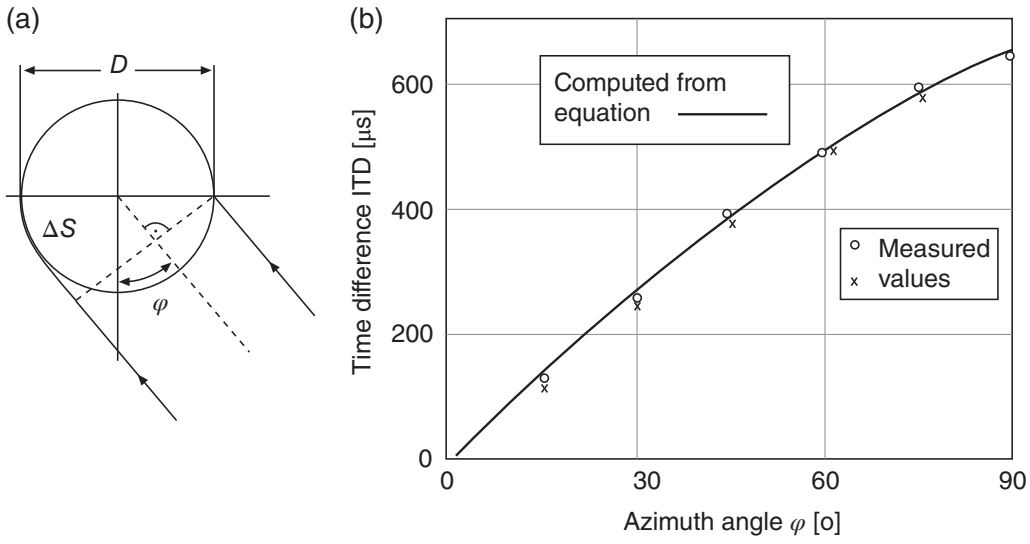

Figure 12.9 (a) The interaural time difference due to propagation path difference ΔS. (b) The ITD in microseconds plotted as a function of azimuth angle φ in the horizontal plane, computed using Equation (12.2) (solid line) and measured values (markers ∘ and ×).

The ITD varies from 0 μs to approximately 600–700 μs when the azimuth angle varies from 0° to 90°. The values derived by the spherical model match the measurements from real heads well, as shown in the figure. Figure 12.10a shows the dependency of the ITD on frequency and on direction as a surface. The values are computed using a cross-correlation type of binaural auditory model. The ITD does not depend significantly on frequency. Only at low frequencies below about 700 Hz are the maxima of the ITDs a bit higher. Some irregularities in ITD functions at frequencies above 2 kHz also exist, and these are due to the complex multi-path wave propagation on the contralateral side. The coherence of the binaural signals is typically lower there, which causes the irregularities in cross-correlation computation.

The hearing mechanisms are sensitive to the ITD between the critical bands of the left and right ears sharing the same frequency. For low-frequency signals up to about 1.5 kHz, the auditory system is sensitive to the phase difference between the narrowband signals in each auditory band. Note that the cue is still often called the ITD, although in some cases the slightly more precise term interaural phase difference (IPD) is used. For frequencies higher than about 800 Hz, the size of the head is comparable to or larger than half the wavelength of the sound, making phase differences ambiguous. Nevertheless, humans are sensitive to the IPD at frequencies up to about 1.6 kHz (Blauert, 1996). ITD cues are still extracted at higher frequencies by the auditory system from the delays between temporal envelopes of the signals.

The perceived lateral position of an auditory event with varied ITD cues while keeping the ILD constant has been measured in headphone experiments, and the results are shown in Figure 12.11a. The position is measured with impulse-like stimuli having a time difference τ_{ph}. In principle, the perceived lateral position shifts in the direction of the ear at which the signal arrives first. For time differences from -600 μs to +600 μs, the dependence on lateral position is linear, and for time differences in the range of about 1–20 ms, the sound is perceived in the ear corresponding to the preceding impulse (Blodgett *et al.*, 1956). With even larger ITDs and for continuous signals, the subjects can no longer tell on which side the auditory event occurs; they may perceive one on each side or perceive a diffuse source in all directions.

Humans are relatively sensitive to small changes in ITD. The value of the JND of the ITD depends on the ITD itself, on the level of the signal, and on the temporal and spectral content of the signal. In the best case, with relatively broadband stimuli which have strong temporal envelopes, and with base ITD corresponding to directions near the median plane, the JND of ITD is of the order of the 10 μs to 20 μs (Blauert, 1996; Hafter and De Maio, 1975).

12.3.2 Interaural Level Difference

The *interaural level difference*, ILD, is another major binaural cue. It is commonly known as the level difference between the ear canal signals. Plane waves produce an ILD due to the scattering – the shadowing and reflection – of sound waves by the head. When a sound wave from a distant source arrives at the head, the pressure level increases due to reflection at the ipsilateral ear and the pressure level decreases at the contralateral ear, as shown in the centre panel in Figure 12.5.

Scattering is a frequency-dependent phenomenon, since wavefronts arriving from distant sources – plane waves – produce ILD only where the head is large enough compared to the wavelength to affect the acoustic field around it. As can be seen from Figure 12.7, the level of the HRTF magnitude responses is constant for frequencies below about 400 Hz, and the dependence on the azimuth angle increases for higher frequencies, reaching more than 20 dB at frequencies above 4 kHz. Therefore, it is a natural consequence that the ILD is a strong

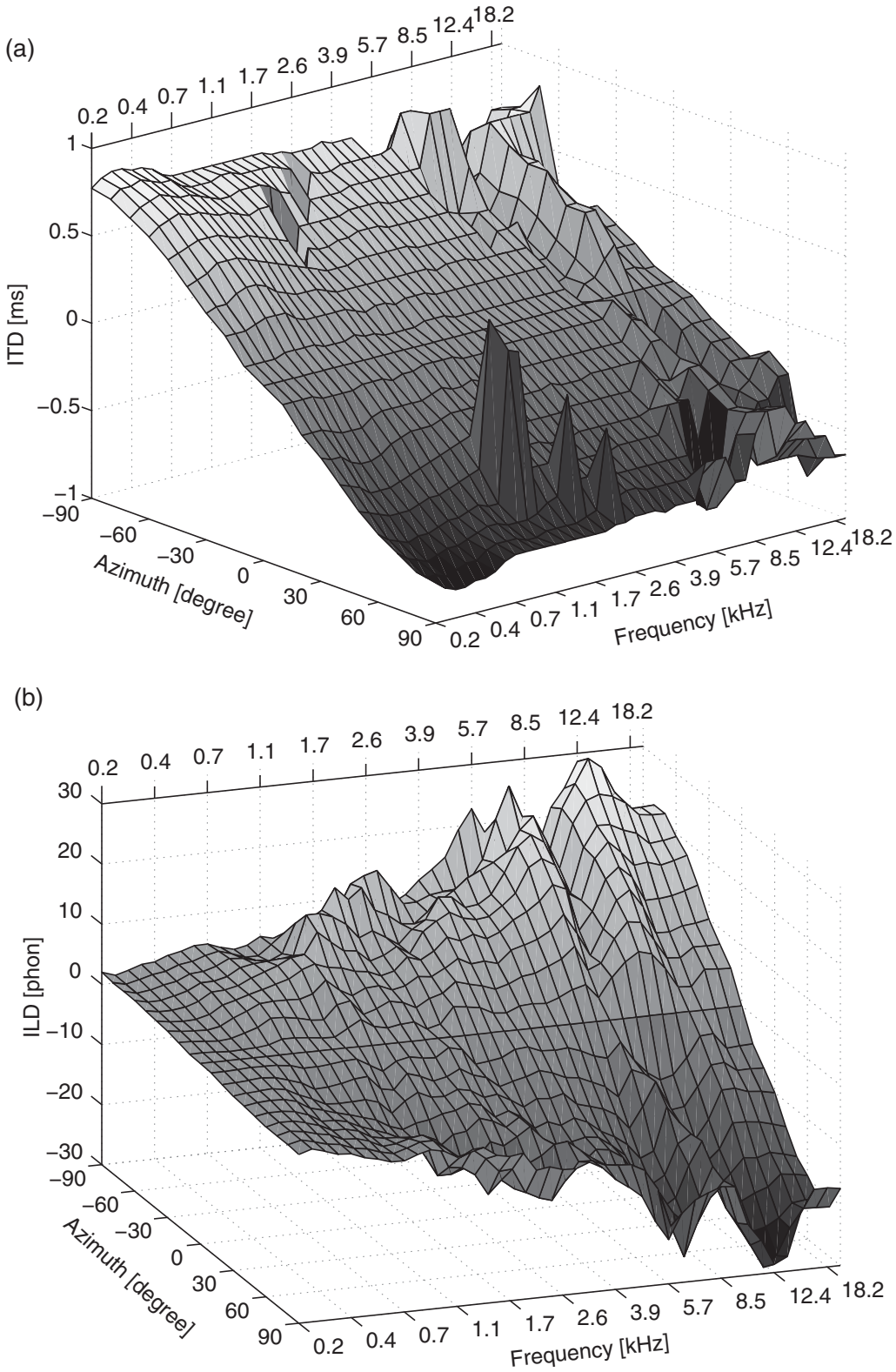

Figure 12.10 (a) The interaural time difference (ITD) functions created by a sound source emitting pink noise in different azimuth directions and in different frequency regions. The ITD is computed using a Jeffress-type auditory model based on the cross-correlation between ear canal signals (see Section 13.5.1). (b) The interaural level difference (ILD) functions expressed as the loudness-level difference in phons computed using a simple auditory model from the same measurement as in a).

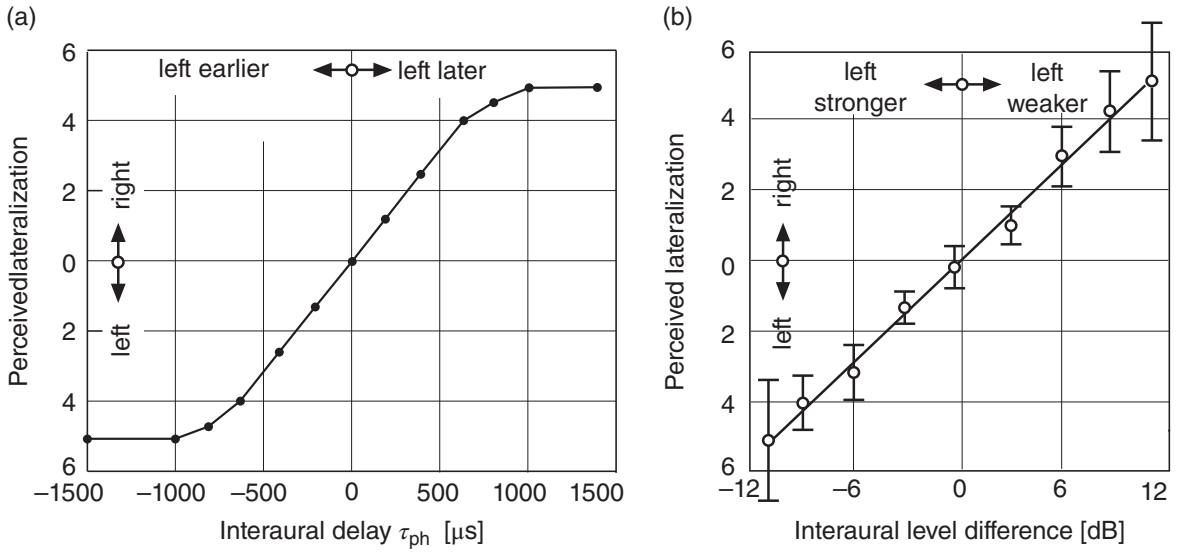

Figure 12.11 (a) The perceived lateralization position as a function of the interaural time difference and (b) the level difference. The ITD experiment is based on impulse-like stimuli (Toole and Sayers, 1965) and the ILD experiment on a 600-Hz sinusoid as the stimulus (Sayers, 1964). Reproduced with permission from The Acoustical Society of America.

directional cue for plane waves (or distant sources) at high frequencies. The magnitude of the ILD as a function of source sound direction and frequency is shown in Figure 12.10b, which shows the frequency dependency clearly.

The sensitivity of lateralization to ILDs using sinusoidal stimuli is shown in Figure 12.11b, which indicates the perceived lateral position of the auditory event created with a 600-Hz sinusoid having an interaural level difference. The level difference range of approximately $-12 \ldots + 12$ dB covers the lateral scale between the ears, so that the auditory event is displaced towards the direction of the louder signal (Sayers, 1964). A complete saturation of the localization of the auditory event in the left or right ear is obtained when the ILD has an absolute value of 15–20 dB.

An apparent inconsistency in spatial hearing is the fact that although plane waves cause an ILD only at high frequencies, humans are sensitive to ILDs at all frequencies. As shown in Figure 12.12, the JND of the ILD is less than 3 dB with sinusoids at frequencies below 2 kHz for absolute ILD values less than 22 dB. With larger absolute values and at higher frequencies, the JND is 3–7 dB. The lowest JNDs are obtained when the ILD value is near zero (Weiping *et al.*, 2010).

As noted earlier, plane waves do not generate ILDs when their wavelengths are long compared to the size of the human head. The question then remains, *why are we sensitive to ILD at low frequencies?* There are at least two reasons for this. At low frequencies, the ILD is clearly a distance cue, since when the source reaches the listener, much higher ILDs are measured between the ear canals. The other use of the ILD seems to be in the perception of coherence of ear canal signals. It has been shown that humans are sensitive to very fast changes of ILD in time; the time resolution of ILD changes is of the order of a few milliseconds. If the ear canal signals are incoherent, as may happen in the case of diffuse sound, large instantaneous fluctuations in ILD are analysed (Goupell and Hartmann, 2007).

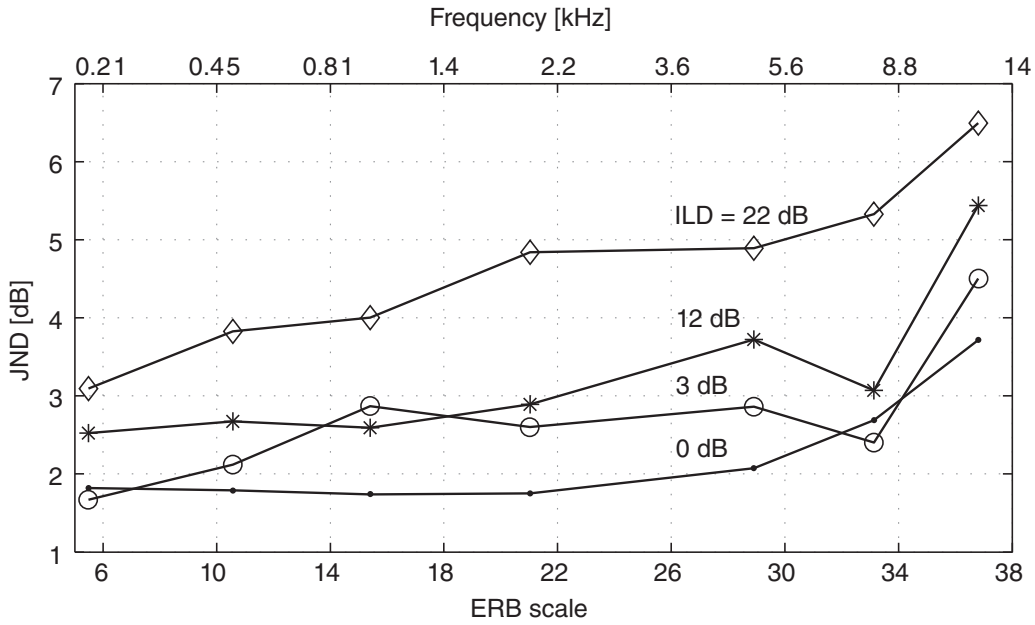

Figure 12.12 The JND of the ILD as a function of frequency and the base ILD. Adapted from Weiping *et al.* (2010)

12.3.3 Interaural Coherence

Interaural coherence (IC) is a measure of the similarity of the signals at the ear canals (Faller and Merimaa, 2004). Humans are sensitive to IC, but it is not clear if the auditory system computes it as an independent cue or if it is perceived due to the effect it has on the values of other directional cues. Some binaural models, which will be discussed further in Section 13.5.1, assume that IC is computed in the auditory system. In this section we define the coherence simply by stating that the coherence of ear canal signals is close to unity when listening to a single sound source in a free field. Low coherence is obtained in a diffuse field, or when sound from multiple sources arrives from different directions at the listener. The value of IC also depends on the frequency in the diffuse field. Low coherence is obtained only at frequencies above about 400 Hz, where the distance between a subject's ears is comparable to the wavelength (Borss and Martin, 2009).

When IC is high, as when in an anechoic chamber, the auditory events are localized to be point-like. When a broadband sound signal, such as pink noise, is presented with low IC, either using headphones or multiple loudspeakers, subjects perceive an auditory event that is located in surrounding directions or in many directions at the same time (Blauert, 1996). It has been found that humans are especially sensitive to coherence at low frequencies with narrowband signals (Culling *et al.*, 2001). However, the perception of broad sources is not limited to low-frequency signals only. When multiple loudspeakers are used to reproduce incoherent high-frequency content, a spatially broad auditory event is perceived (Santala and Pulkki, 2011).

In normal rooms, the IC is high when the onset of an impulsive sound arrives at the ear canals, and the IC is lower when reflections and the room effect in general arrive at the ear. Faller and Merimaa (2004) suggest that localization occurs only if the IC has a high enough value. Alternatively, as discussed in the previous section, low coherence also causes ILD and ITD cues to fluctuate randomly with time and frequency (Goupell and Hartmann, 2007). These

fluctuations, caused by the low IC, are assumed to lead directly to a perception of surrounding auditory events, and no separate IC cue is computed in the auditory system.

12.3.4 Cues to Resolve the Direction on the Cone of Confusion

At least in principle, the ITD and ILD do not change when changing the position of a sound source on a cone of confusion. This means that decoded values of the ITD and the ILD do not resolve the direction from which the sound arrives at the listener, since the same cue values are produced by sound arriving from any direction specified by the surface of the cone. The fact that we can perceive the front–back dimension quite well and the elevation angle correctly indicates that there must be effective mechanisms for resolving the direction inside the cones of confusion. There are two mechanisms by which we perceive the elevation correctly, the analysis of spectral cues and the utilization of dynamic cues, which will now be discussed.

Spectral cues

Monaural cues, based on information from a single ear or common to both ears, are important in localization. Because no interaural differences are used, the information for monaural cues is derived from the ear canal signal itself, particularly from the properties of the sound source and its variation due to the effect of the head, but also due to the acoustic environment. The main monaural cues, when the direct sound from a source is dominant, are *spectral cues*. Since the monaural time resolution of hearing is about 1–2 ms, and as most of the details in HRIR are located within a 1-ms time window, monaural temporal cues are considered unimportant.

Due to scattering and reflection caused by the head and the pinna, the magnitude spectrum of the sound entering the ear(s) is dependent on the direction of arrival. These cues are denoted as *spectral cues*. Spectral cues are important when localizing sound sources in the median plane and its vicinity. To illustrate the spectral cues, the left ear HRTFs measured with a source in different directions in the median plane are shown in Figure 12.13 for four subjects. It can be seen that the HRTFs indeed carry information on direction; there are vast changes in the spectra when the elevation changes. The pinnae of subjects are different, and so are the spectral cues. The figures have similar structures, the arcs have similar shape, but they are located at different frequencies. The notches and humps at 4–10 kHz are also located at different frequency-elevation positions, although some similarities between subjects are evident. The overall black–white contrast in the plots is also different. For example, subject 17 has less contrast than subject 12, which implies that the spectral notches are deeper with subject 12 than with subject 17.

Broadband signals are generally needed to make spectral cues efficient, since otherwise the cues cannot be decoded. This is a fair assumption, since most natural sounds have a relatively flat spectrum at frequencies where the pinna cues are effective, namely at 4–10 kHz. On the other hand, it has been shown that when a sinusoid is presented in that frequency region, using a static loudspeaker in the median plane, the perceived direction depends significantly on the frequency of the sinusoid. The listener perceives the auditory event as moving in space when the frequency is changed (Blauert, 1996), although the sound source does not move.

Learning also has an effect on the utilization of spectral cues. This is natural, since the ears of humans develop steadily throughout life, and the hearing system has to continuously adapt to the acoustic effect of the pinnae. In an experiment by Hofman *et al.* (1998), the cavities of subjects were made smaller using putty, thus destroying the natural spectral cues. The subjects

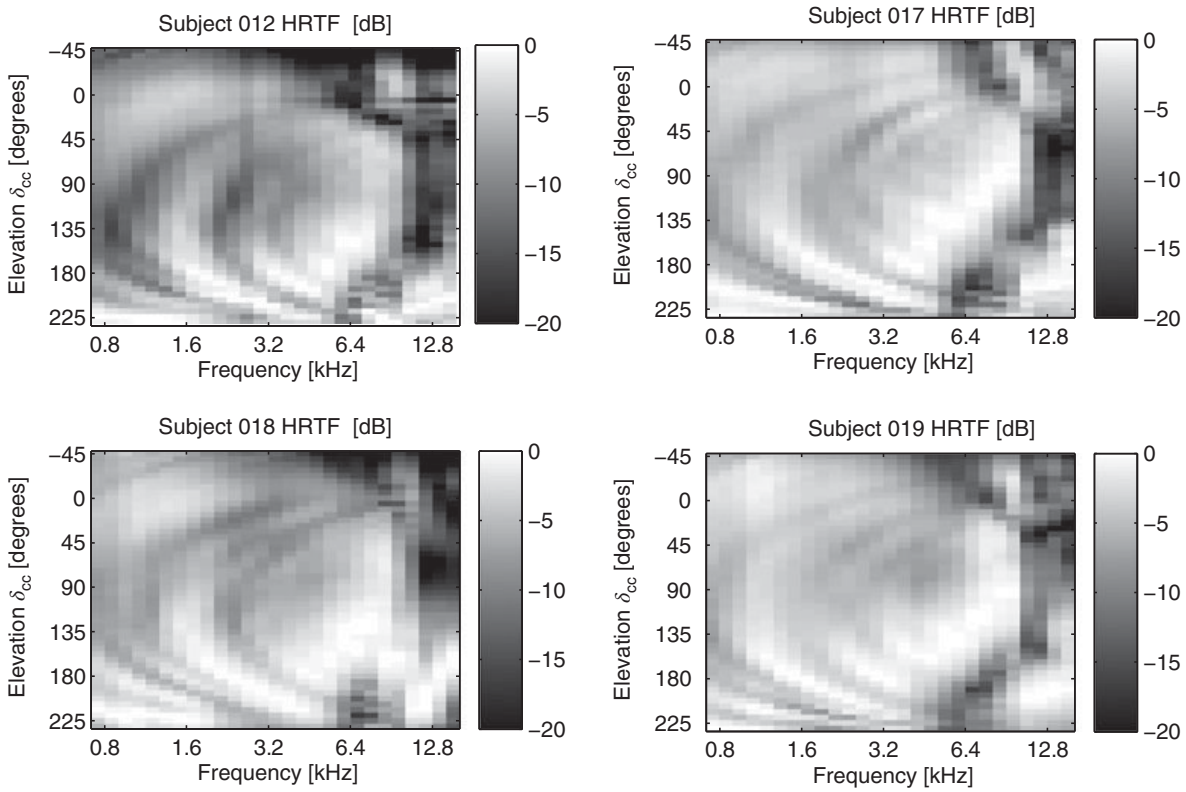

Figure 12.13 The HRTF magnitude spectra measured for 50 elevations in the median plane and smoothed over 1/6th-octave bands. The smoothed magnitudes have been normalized with the maximum of the magnitude in each frequency band. Different panels show the results of different subjects. The HRTF data for the illustrations originate from Algazi *et al.* (2001).

immediately lost their ability to perceive elevation, however this returned in a few weeks to normal if the putty was not removed from the ears during that time. The listeners thus learned the acoustic properties of the new ears during the period they had putty in their ears. Very interestingly, after removing the putty, the listeners regained their elevation perception ability immediately. For some reason, no after-effect was present.

Another case where monaural (spectral) cues are of the utmost importance is when a listener has only one ear in effective use. Localization of sound sources is still possible, although the accuracy is strongly degraded (Blauert, 1996).

Dynamic cues

The other mechanism humans use in resolving the direction inside the cone of confusion is the utilization of dynamic cues (Blauert, 1996). When the listener moves his or her head by rotating, tilting, or moving, the binaural cues change accordingly. Let us assume that a sound source is in front of the listener. If the listener rotates his or her head right, the left ear comes closer to the source and the right ear moves farther away. This changes the ITD and ILD, favouring the left ear. If the source was located behind the listener, with the same movement the binaural cues would change to favouring the right ear. The localization mechanisms implicitly assume that the sound source locations are static in the global coordinates, and thus the sources do not move in synchrony with the listener's head movements, which most often is true. Thus,

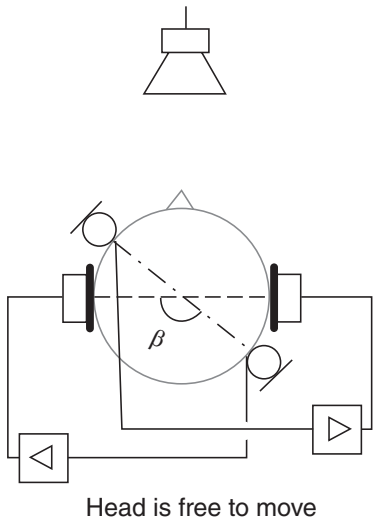

Figure 12.14 A set-up for pseudophone localization experiments. Microphones away from the listener's ears move with head rotation.

the dynamic changes of the ITD and the ILD during head movement provide information on the position of the source on the cone of confusion.

An interesting experimental set-up called *pseudophone localization* is presented in Figure 12.14. Signals from microphones are reproduced through headphones so that the subject perceives sounds as if having ears at the positions of the microphones. A particularly confusing situation is achieved if the angle β is 180°, so that the ears are effectively interchanged, resulting frontal sound sources generating auditory events behind the listener, and vice versa (Blauert, 1996). One of us (Ville Pulkki) had a striking experience with such a device. A person was talking in front of me at a distance of about one metre. When I rotated my head with the pseudophone in action, the auditory object jumped immediately to behind me. Although the talking mouth of the person was clearly visible in front, the auditory object remained behind. The dynamic cues thus overrode the visual cues in this case.

One of the reasons why headphone reproduction of audio is, in most cases, internalized is the lack of dynamic cues. In such cases, the binaural cues in audio reproduction do not change at all when the listener moves his or her head. The only situation where cues remain constant when the head is moved is when the source is inside the head, and the localization mechanisms take this as strong evidence that the sources are indeed inside the skull. Note that localizing sounds inside one's own head is not at all unusual. Typical inside-the-head-localized sounds are, for example, one's own voice and eating and breathing sounds.

The dynamic cues thus seem to provide strong evidence as to whether the source is in front, behind, up, down, or inside the head. This effect is of great interest in headphone-based binaural reproduction of sound, where the position of the head is updated in the reproduction system. As will be discussed in Section 14.6.2, the correct reproduction of dynamic ITD and ILD cues mitigates directional errors perceived due to erroneous monaural cues. This underlines the effectiveness of dynamic cues.

12.3.5 Interaction Between Spatial Hearing and Vision

Hearing is only one of the sensory mechanisms for communication and receiving information from the environment. Spatial perception and sound-event localization may be influenced by

non-acoustic cues as well, particularly by visual cues. In *multimodal perception*, the peripheral sensory mechanisms work relatively independently, but these partial percepts are fused together into a coherent internal representation, unless conflicting cues leave them apart or the dominating sensory organ overwhelms the weaker evidence.

A well-known example where vision modifies the auditory perception is ventriloquism (Alais and Burr, 2004), where listeners perceive the sound to be coming from a direction other than the one it actually comes from, if the visual movements associated with the sound generation are synchronized with the sound. The synchronized movements then 'capture' the spatial perception of sound. The effect is most powerful if the separation between the sound and visual sources is less than about 30°. However, the auditory cue may still be more salient than the visual cue in some cases. If the visual image is severely blurred, the perceived direction matches that of the sound source, and if both of them are blurred, an average direction is perceived.

12.4 Localization Accuracy

The previous sections described the cues available to human listeners to localize a sound source. This section discusses the accuracy of humans in terms of localization. However, the discussion is limited to listening to single real sources in a free field. Localization accuracy can be much worse in spaces with reflections and/or reverberation, as will be discussed in subsequent sections.

The signal content also has a strong effect on localization accuracy and localizability of sound sources. Accuracy is best with broadband sounds with strong temporal fluctuations, such as transients, speech, and noise in general. On the other hand, narrowband sounds, such as mosquito or certain bird sounds, can be very hard to localize, especially in rooms. This can be understood from the perspective of evolution. The correct localization of predators or prey is vitally important for survival, and the sounds generated by movements are typically broadband, impulse-like sounds.

12.4.1 Localization in the Horizontal Plane

In the best case, with broadband sounds in a free field, our directional hearing locates sound sources accurately. Figure 12.15 shows the results from an experiment where the perceived direction of a white noise sound source was estimated by subjects in four main directions (front, left, right, back). The sources on the left and right sides were placed approximately 10° towards the frontal direction. Localization blur, or the variance in the perceived angle, was lowest at the front, somewhat bigger at the back, and largest in the side directions.

One method to evaluate the localization resolution is to measure the JND of direction for different frequencies of tones. Figure 12.16 plots the results of such an experiment using sine-wave test signals (Stevens and Newman, 1936). The best resolution is about 1° for tones below 1 kHz coming from the front of the subject, and it deteriorates towards the sides. At around 1.7 kHz the resolution is the worst, especially for sounds from the sides, so much so that the subjects cannot make any reliable estimate at all. This is the frequency where the ILD does not depend monotonically on direction and the head size is about one wavelength, leading to maximal confusion in the ITD and ILD decoding. At higher frequencies, the resolution improves, but is never as good as for frequencies below 1 kHz. Note that the JND values do not reflect the localization accuracy, but only how much a source has to be moved for the change to be noticed.

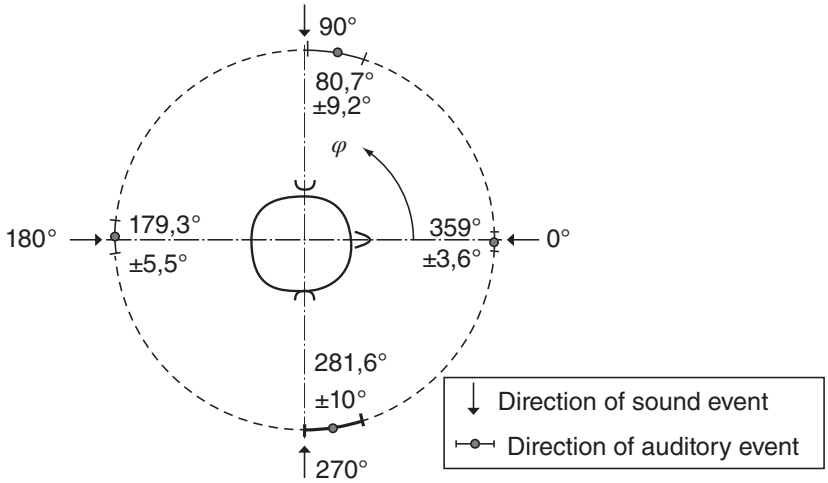

Figure 12.15 The perceived azimuth angle in the horizontal plane in free-field conditions for a white noise sound source in four primary directions (0°, 90°, 180°, and 270°). The average perceived angle and the localization blur are marked by a thick line. Adapted from Blauert (1996).

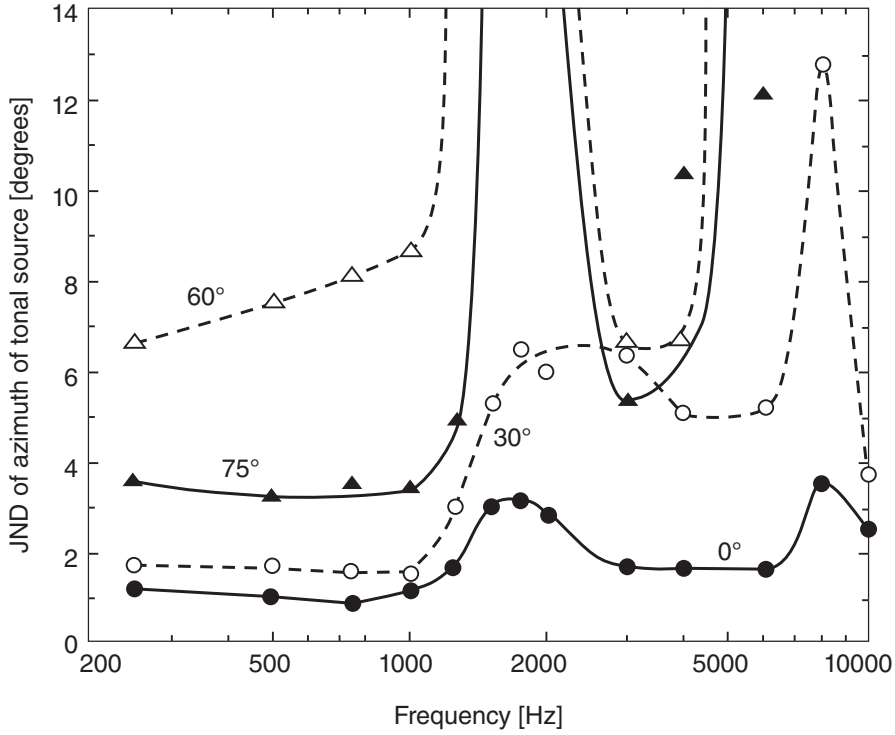

Figure 12.16 JND in the azimuth angle ($\varphi = 0°, 30°, 60°, 75°$) of a sound source emitting a tone in the horizontal plane as a function of frequency. Adapted from Mills (1958).

12.4.2 Localization in the Median Plane

The accuracy of perceived elevation of a sound event in the median plane is worse than the corresponding accuracy of perception of azimuth directions, especially with an immobilized head. Experimental results are shown in Figure 12.17 for a set of elevations using speech as the test signal. There is a noticeable bias to the frontal direction, and the localization blur increases for sources above and behind the subject. The lower accuracy compared to azimuth perception

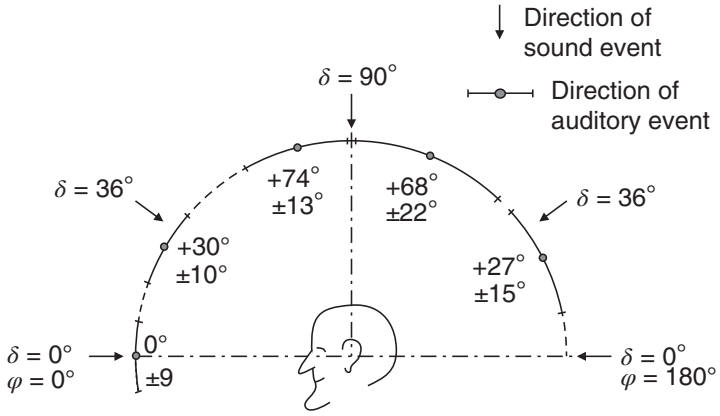

Figure 12.17 The perceived elevation and localization blur in the median plane for five physical elevation angles ($\delta = 0°, 36°, 90°$ frontally, and $0°, 36°$ from the back) with speech as the test signal. Adapted from Damaske and Wagener (1969).

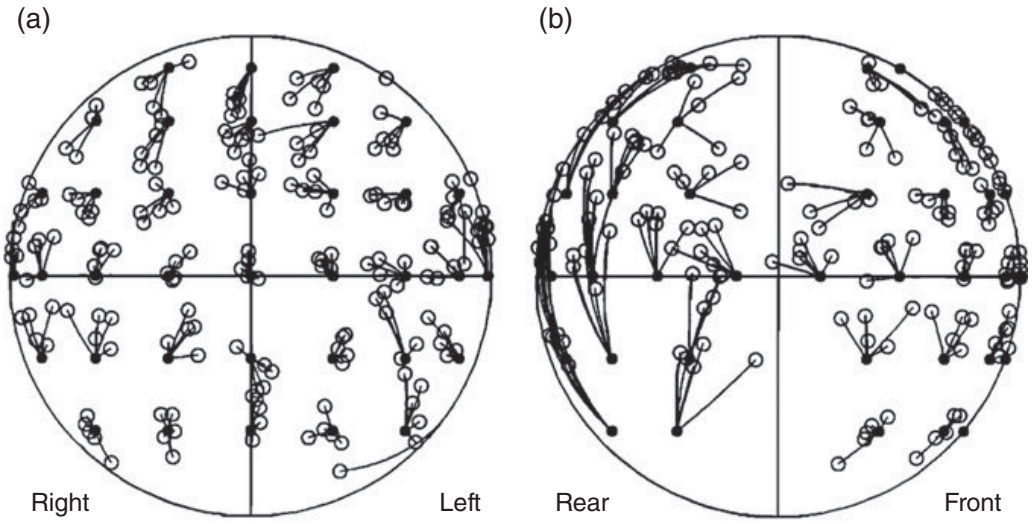

Figure 12.18 Localization results of 250-ms broadband sound sources in a low-echoic room with a spacing of 20° in azimuth and elevation. The filled dots show the positions of the sound sources and circles the responses of the same subject to five presentations of sound. The task of the listeners was to point towards the source with their nose. (a) Projection of the results in front of the listener. (b) Projection of the results to the front and rear of the listener. Adapted from Middlebrooks (1992). Reprinted with permission from The Acoustical Society of America.

is due to the fact that the ITD and the ILD cues cannot be utilized in the median plane, and spectral cues are the main source of information for the immobilized listener. If subjects are allowed to rotate or tilt their head, the change in ITD and ILD cues provides more information on the direction of the source.

12.4.3 3D Localization

A number of studies have been conducted to measure the accuracy of directional hearing when any direction in three dimensions is allowed. Figure 12.18 shows a result from a test where 250-ms noise bursts from different directions were presented to subjects, and their task was to point with their nose to the direction of the sound source (Middlebrooks, 1992). The results

in Figure 12.18a show quite consistent accuracy within a few degrees for frontal directions, especially near the horizontal plane. Slightly larger errors occur when the source departs from the horizontal plane. Figure 12.18b shows the performance in the rear directions. It is clear that the subjects overestimate the elevation of the sources at the back, a result that is in agreement with the median plane localization discussed in the previous section.

The results show that the accuracy of perception of elevation degrades in directions outside the field of vision. This is in line with the fact that subjects continuously adapt to their spectral cues, as discussed in Section 12.3.4. The result also suggests that adaptation can occur only if the sources are visible. Using vision as the reference for adaptation of directional hearing is a viable option, as vision is the most accurate method to measure the direction of a sound source for adaptation. Furthermore, the auditory and visual pathways meet at a relatively low level in the brain, as discussed in Section 7.7.1, which could facilitate such adaptation.

The results on perceived direction and localization blur vary in different experiments with the test signals used, the test subjects and their familiarity with the task, the method of registration of the perceived direction, and so on. Thus, the values given in Figures 12.15 to 12.18 characterize only approximately the behaviour of directional hearing.

12.4.4 Perception of the Distribution of a Spatially Extended Source

White water, or waves hitting the seashore, forms a wide sound source from the perspective of an observer. A listener, indeed, perceives such sound to be wide. The ability to perceive the spatial distribution of a spatially extended sound source is not very accurate, especially if the type of sound signal is not optimal. Humans are at their best in this task when the sound from different sources is noise-like or impulsive, or if different parts of the source emit different frequency content. On the other hand, if all the parts of the source emit sinusoids with equal frequency, the subject perceives a very narrow auditory event that does not depend on the actual width of the source.

Figure 12.19 shows some results from a listening experiment where different subsets of 13 loudspeakers in the horizontal plane in an anechoic chamber were used to emit mutually incoherent pink noise at equal levels to the subject. The task of the subject was to indicate which loudspeakers emitted the sound. The loudspeakers which actually emitted the sounds are marked on the figure with black squares, and the grey bars show the percentage of 'sound on' indicated for each loudspeaker by the ten subjects for two repetitions of the task. The three topmost panels show that the subjects indicated the spatial distribution correctly only in the case when the number of sources was one, two, or three (Santala and Pulkki, 2011). For a wide and dense constellation of sources, as in the fourth panel, the edges of the distribution were perceived almost correctly, although a bit biased towards the centre. However, the perceived distribution in the central area of the source did not match at all with the actual distribution of the sound source. The two panels at the bottom show spatially complex scenarios, where the subjects clearly had no clue which loudspeakers were actually on.

The accuracy of the hearing system is thus clearly not at its best when analysing the spatial distribution of sources. Although the directional separation between loudspeakers was 15° or more, the listeners failed to report the distribution accurately. The task of perceiving the distribution in such a case could be thought to be a simple one, since the accuracy of direction perception for a single source is of the order of a few degrees at best. However, the distribution is analysed from signals from the sources summed in the ear canals, and naturally the presence of multiple temporally and spatially overlapping source signals arriving from different directions makes the analysis task difficult.

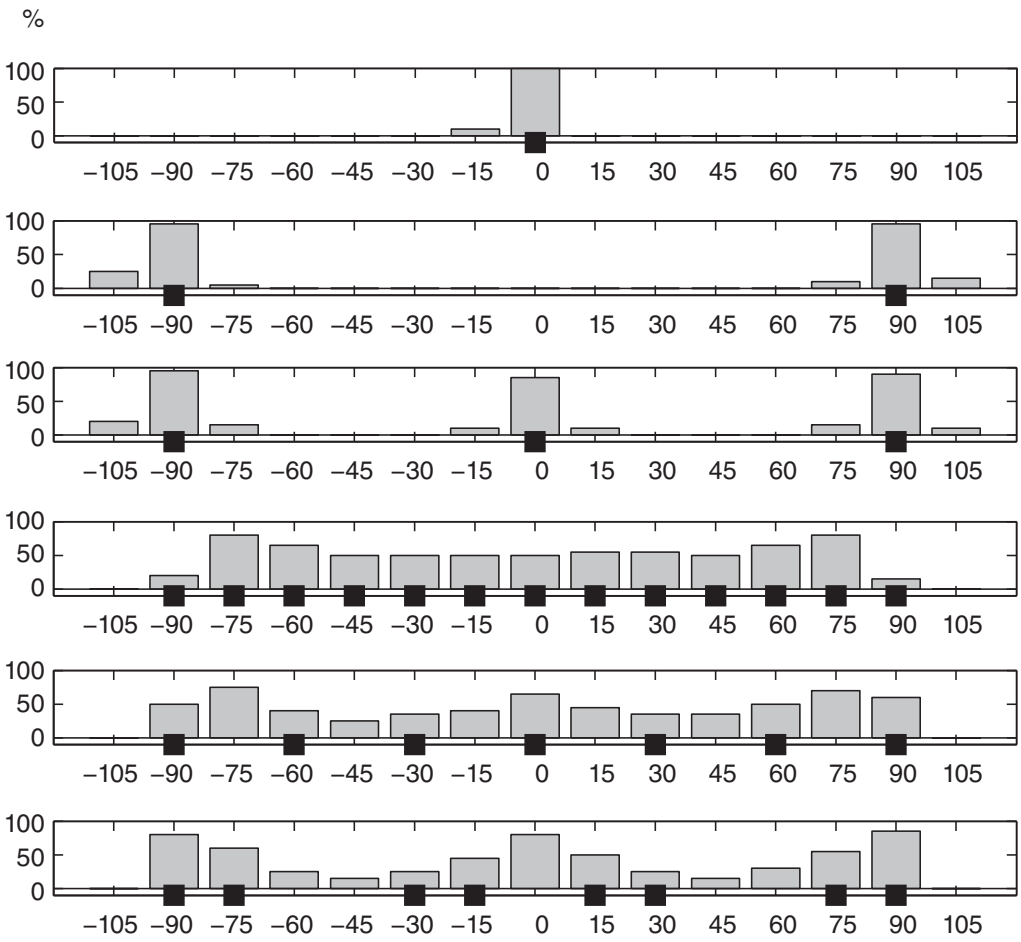

Figure 12.19 The results of a subjective test with 13 loudspeakers set up horizontally in front of the listener in an anechoic chamber. The loudspeakers, marked with black squares, emitted pink noise, and the task of the listeners was to indicate which loudspeakers were actually emitting sound. The grey bars denote the relative rate at which a specific loudspeaker was indicated to emit sound. The azimuth directions of the loudspeakers are represented on the abscissa. Courtesy of Olli Santala.

12.5 Directional Hearing in Enclosed Spaces

Spatial hearing is at its best in situations with point-like broadband sources in a free field, as shown in the previous section. Also of interest is the spatial resolution when strong reflections or reverberation exist in the environment. Our hearing has adapted to such cases with remarkable resilience. The directions of sound sources are perceived correctly in many challenging conditions. The ability of humans to perceive the geometry of an enclosed space, however, is quite limited, but on the other hand, the listener can compensate, at least partly, for the colouration caused by the room response in the detection of timbre.

12.5.1 Precedence Effect

The precedence effect is an assisting mechanism of spatial hearing (Litovsky *et al.*, 1999). It suppresses the effect of early reflections in source direction perception. This helps to localize sound sources in reverberant rooms. The reflections reaching the listener in a reverberant room are added to the direct sound in the ear canals, which changes the binaural cues considerably. Therefore, the only reliable cues are those generated by the direct sound without the presence of reflected sound in the ear canals. In enclosed spaces, direct sound is dominant in the ear

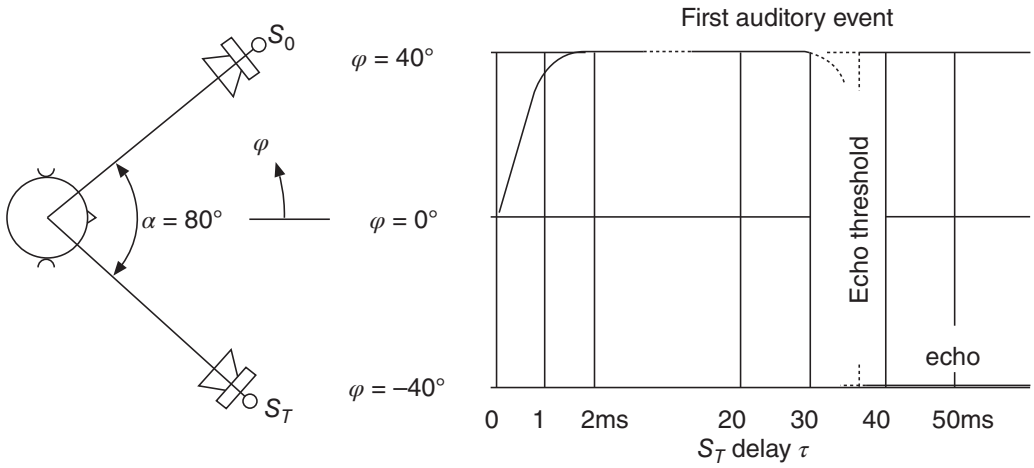

Figure 12.20 The perceived direction of an auditory event as a function of the delay between two impulsive sound events generated with different loudspeakers. Adapted from Blauert (1996), and reprinted with permission from MIT Press.

canals only shortly after the onset of sound after silence, before the reflections arrive at the listener. The precedence effect takes advantage of this.

The typical research scenario is shown in Figure 12.20, where two loudspeakers are placed in the azimuth directions of $\pm 40°$ in anechoic listening conditions. The *lead* sound event S_0, typically with an impulsive sound, is presented at time instant 0 ms by the loudspeaker at 40°, and an identical signal is presented as a *lag* sound event, S_T, by the loudspeaker at $-40°$. The delay τ between the sound events is varied. When τ is zero, the listener perceives a single sound event in front, and when τ is increased to about 1 ms, the single spatial auditory event migrates towards the lead sound event. The spatial perception remains the same until the onset of the echo threshold, which occurs for values of τ of about 30–40 ms, depending on the signal. Only after this threshold is crossed does the lag sound event create an additive, spatially separated auditory event.

If a level difference is introduced between the lead and lag sound events, different thresholds are found in the precedence effect. As shown in Figure 12.21, the louder the lag sound is made compared to the lead, the more probable it is that it will be audible. The thresholds also depend on the delay and, in principle, the larger the delay is, the more prominent the lag. The lowest threshold is the detection threshold of the lag. When the level of the lag sound exceeds this threshold, its presence is noticed as a difference in timbre; it does not produce a spatially separated auditory event.

Evidently, our hearing responds strongly to onsets and transients, and the binaural cues decoded from the short response with a length of only about 1 ms have, correspondingly, a strong emphasis. After a strong response from most neurons, the refractory period makes the directional hearing mechanisms non-functional. Thus, the direction of the delayed sound does not affect the perception.

12.5.2 Adaptation to the Room Effect in Localization

As already mentioned, the localization of sources is based on signal analysis in the ear canals, and reflections and reverberation largely ruin the monaural and interaural cues. The precedence

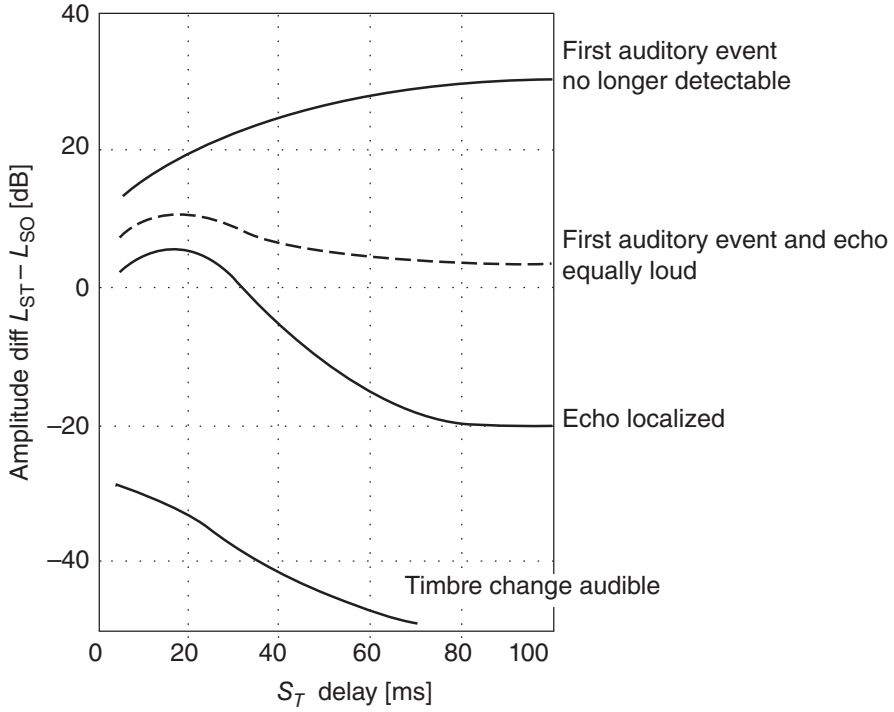

Figure 12.21 Different thresholds in the precedence effect scenario as a function of delay and level difference between sound events at time instants 0 ms and T. Adapted from Blauert (1996).

effect to be shows that transient information is judged most relevant in the detection of the left or right direction. In addition to the precedence effect, there are also other mechanisms aiding the localization tasks that are not known, but their existence has been shown by the fact that the accuracy of subjects' performance in localization tasks in rooms constantly improves with the number of trials, even after repeating several hundreds of trials (Shinn-Cunningham, 2000; Zahorik *et al.*, 2005). The subject gathers some data on the spatial cues generated by the room and somehow adapts to them, perhaps by using templates.

12.6 Binaural Advantages in Timbre Perception

A relevant question is how much the binaural hearing aids in decoding the sound spectrum emanated by multiple spatially separated sources, also in reverberant spaces. In vision, the spatial separation of sources definitely helps in the process, because when visual stripes are separated by more than about 0.1°, they are discriminated. The working principle of the eye and the ear, however, differ greatly. The eye has a lens and a multitude of receptors primarily sensitive to the different directions from which light comes, while in contrast the ear has only one cochlea. Spatial hearing is conducted as an analysis of the ear canal signals, and no primary sensitivity to direction exists in the cochlea. A lesser advantage is expected with hearing, but, as will be shown below, a significant benefit is still obtained.

12.6.1 Binaural Detection and Unmasking

How well humans can separate the signals emanated from spatially separated sources is an interesting subject. Since we have two ears, we should have some capability to listen selectively

to different directions. Such a release of masking due to binaural hearing is called *binaural unmasking*.

Interestingly, humans are capable of detecting the content of sound coming from different directions in the presence of distracting sources. The speech signal has been widely used as a the test signal in a the study of this capability, as it presents a natural signal to which human hearing has adapted. In the studies, a speech signal S with masking noise(s) N is presented to the listener. The level of the signal is measured for both binaural listening and monaural reference listening so as to produce the same intelligibility in both cases, and the difference in level between the monaural and binaural conditions is taken as the *binaural intelligibility level difference* (BILD) (Blauert, 1996).

Figure 12.22a shows the BILD in the case when a broadband, speech-like masker noise is presented via six loudspeakers around the subject. The masking thresholds are measured for different azimuth directions of the signal source. When the measured thresholds are subtracted from the corresponding threshold of the signal in the direction of 180°, the plotted level differences are obtained. It can be seen that with binaural listening the advantage is of the order 5–7 dB compared to the reference case with the loudspeaker behind. When the same test is conducted with single-ear listening, the monaural intelligibility level difference (MILD) plot shown in the figure is obtained. It is remarkable that the difference between the MILD and the BILD is small when the sound source is ipsilateral to the unblocked ear. This leads to the

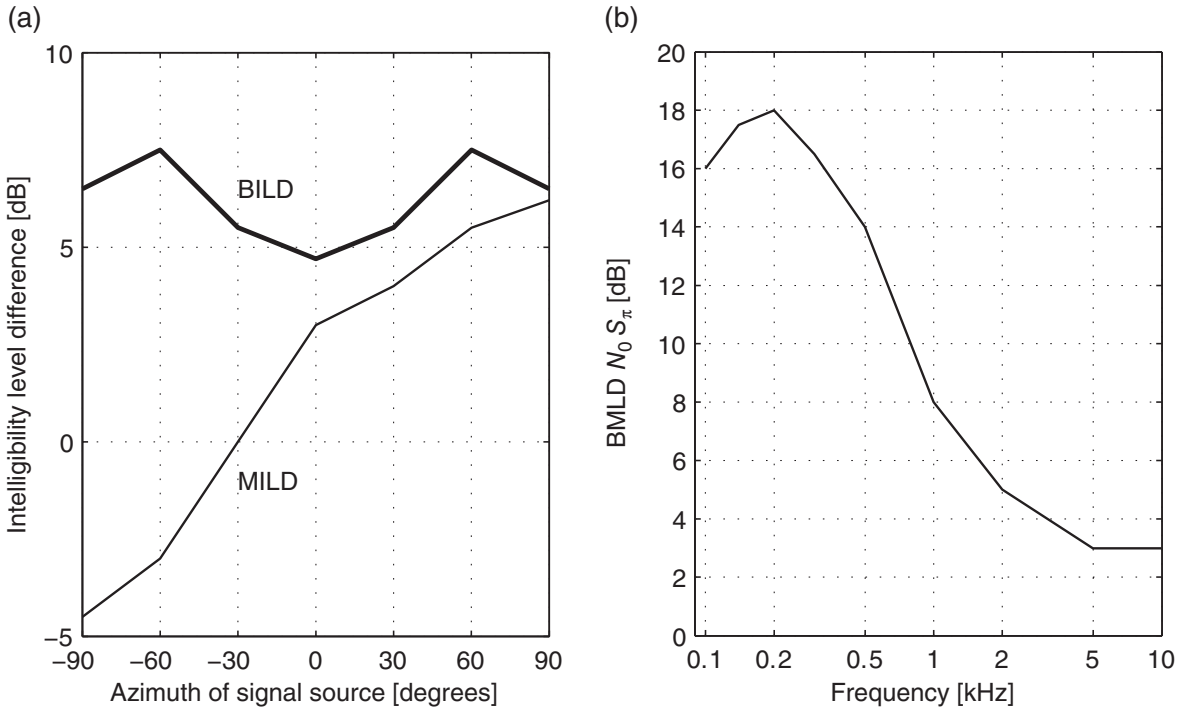

Figure 12.22 (a) Static broadband noise is presented to a subject through six surrounding loudspeakers. The level resulting in a certain degree of intelligibility has been measured using binaural and monaural listening. The direction of the signal source is shown on the abscissa and the BILD and the MILD on the ordinate. The level differences are computed in the reference case using a signal in the direction of 180° with monaural listening. (b) The binaural masking level difference with a broadband diotic masker and an antiphasic sinusoidal signal depending on frequency. Figures adapted from Blauert (1996) and reprinted with permission from MIT Press.

assumption that in such cases the signal detection relies on the 'better ear'; the signal is simply decoded from the ear canal signal that has a higher SNR. There is evidence, however, that subjects utilize instantaneous ILD to further improve the intelligibility from 'better ear' listening (Bronkhorst, 2000) by a few decibels.

A related phenomenon is the *cocktail party effect* already mentioned in Section 11.7.1, where the subjects have to follow a speech source among a number of distracting speech sources (Bronkhorst, 2000). In multi-talker conditions, the advantage provided by binaural hearing has been found to be 0–8 dB, which is in agreement with the results of BILD tests.

Certain synthetic signals reveal the significantly greater advantages of binaural listening. The binaural masking level difference (BMLD) has been researched extensively with headphone listening. The most dramatic effect is obtained when the subject is presented with a broadband diotic noise (N_0) and binaurally antiphasic sinusoidal signal S_π. The reference case is $S_\mathrm{m}N_\mathrm{m}$, where both the signal and noise are presented monotically. The BMLD is plotted as a function of frequency in Figure 12.22b, where it can be seen that when the frequency of the signal is about 200–300 Hz, the binaural advantage in signal detection is of the order of 18 dB (Blauert, 1996). Although such a big advantage is not obtained with distant sound sources, the result is interesting in the study of hearing.

Our relatively low capability to listen selectively to different directions may reveal something about the role of hearing in our evolution. Vision is used to precisely locate an object, and when looking in one direction, the perception of visual details of objects in other directions, is greatly diminished, and objects outside the field of vision are not perceived at all. We can assume that evolution has developed hearing as an omnidirectional alarm system with high sensitivity over a large frequency region to complement directionally selective vision. We can only guess at the capabilities of the auditory system if our hearing had developed towards better directional sensitivity. We would probably have more than two ears with poorer frequency resolution, or perhaps acoustic lenses coupled with more of ears.

12.6.2 Binaural Decolouration

Listeners are also able to adapt to spectral colouring caused by the room using binaural listening, which is called *binaural decolouration*. The goal of the mechanisms is interpreted to be the estimation of sound signals emanated by sound sources in an enclosed space. They show a remarkable ability to estimate the sound signal emanated by the source by ignoring the spectral contribution of the room response present in the ear canal signals (Brüggen, 2001). The same effect has been discussed in the context of sound reproduction (Toole, 2006). The decolouration effects are both monaural and binaural. Bilsen (1977) states that some pitch phenomena can be explained by assuming that listeners somehow up the spectra of the left and right ears to form a *central spectrum*. Such a summation would make the dips in the room responses less audible, since the dips typically occur at different positions in frequency, which could explain, at least partly, the decolouration effects.

12.7 Perception of Source Distance

The perception of distance is also an important part of the spatial hearing mechanism. It is a multi-faceted mechanism that is not perfectly understood yet (Zahorik *et al.*, 2005). Evidently, the auditory system uses different cues along with knowledge of the acoustic surroundings. Both monaural and binaural cues are used in the task.

12.7.1 Cues for Distance Perception

There are at least four cues used to extract the auditory distance:

- *Loudness*: The louder the auditory event is, the closer it is. Better distance estimates are extracted if an internal reference of the sound is used for comparison. In other words, familiar sounds aid the perception of distance.
- *Effect of room reflections and reverberation*: The more the room affects the sound, the farther away the source must be.
- *Spectral content*: The fewer high frequencies are in the auditory event, the farther away the source must be.
- *Binaural cues*: When the source is closer than about 1 m to the subject, the ILD between the ear canals grows larger, which is used as a distance cue.

The first of the mentioned distance cues, loudness, is naturally the most related to the SPL generated by the source at the listening position. This cue has been studied in an anechoic chamber by placing a real speaker at different distances and asking five subjects to estimate the distance. In this environment, the loudness with which the speaker talks is the only distance cue available to the subject. The perceived distance asymptotically approaches a finite distance less than 10 m, which can be called the *acoustic horizon*, as the physical distance increases, as seen in Figure 12.23.

The second distance cue in the list above is the *effect of room reflections and reverberation*. In everyday acoustic environments, the signals in the ear canals of a listener consist of direct sound and also a considerable amount of reflections from walls or obstacles and reverberation due to repeated reflections. In such conditions, can take advantage of the room cues and is typically more accurate than in free-field conditions (Nielsen, 1993; Zahorik *et al.*, 2005). Research on psychoacoustics often mentions the direct-to-reverberant ratio as a cue for distance perception. The more there is reverberant sound energy compared to direct sound energy, the farther away the listener perceives the source to be.

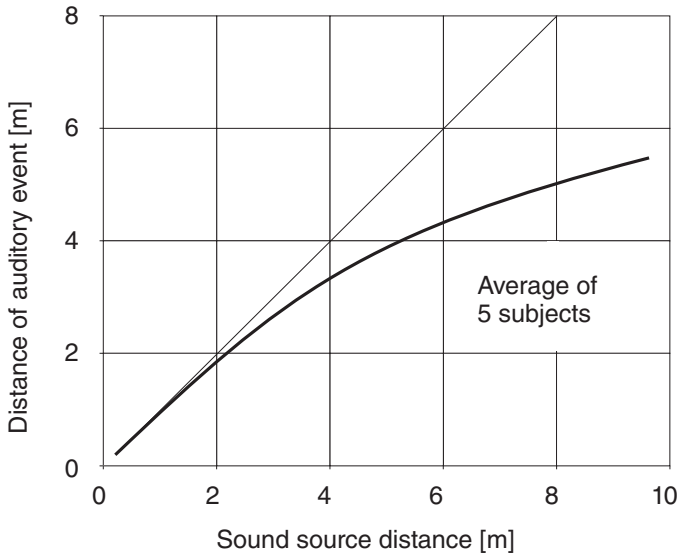

Figure 12.23 The dependency of perceived distance of a real speaker on the physical distance in free-field conditions. Adapted from von Békésy G 1949.

Added to this, audio engineers utilize another method to control distance perception using reverberators, wherein the auditory events caused by at least some types of sounds (typically voice) can be moved closer or farther away by changing the pre-delay parameter. Reverberators are audio effects, which, in principle, convolve a monophonic sound signal with the room response. The pre-delay parameter controls the temporal gap between the direct sound and the rest of the response. Shortening the pre-delay moves the perceived source farther away and making it longer brings it closer. Physically this makes sense, since a long gap would correspond to a situation where the direct path is much shorter than the reflected path, which can only happen if the source is close. This is different from the direct-to-reverberant ratio cue, since the ratio does not change. A very similar effect has been found in virtual reality simulations. When the first reflections are moved temporally in an impulse response without changing the properties of the late reverberation, the perceived distance changes systematically (Pellegrini, 2002).

The third distance cue mentioned in the list is the *spectral content* of sound. It changes when the sound travels long distances in the atmosphere. For example, the sound of lightning sounds like a 'crack' when the bolt is relatively close, and when heard from a distance of several kilometres it sounds more like 'rumbling'. This change is simply due to the fact that the air absorbs high frequencies more than low frequencies. The effect is quite mild, about 3 dB/100 m at 4 kHz. Thus, the listener may use this cue to detect the distances of sources that are far away.

The fourth distance cue in the list is binaural cues, which are essentially directional cues. However, if a source is brought near to a listener from afar in a constant direction, the ITD and ILD cues remain practically the same when the distance r is more than about 1 m. At smaller distances, the ILD cue depends significantly on the distance as well. This is explained by the distance attenuation of sound, which is proportional to the inverse of distance, $1/r$, as defined in Equation (2.21). At larger distances, the $1/r$-law attenuates sound by practically the same amount for both ears, since the distance between ears is negligible when compared to the distance of the source. With shorter distances, the sound entering the farther ear is attenuated more than the sound entering the closer ear due to the distance difference (Duda and Martens, 1998; Shinn-Cunningham *et al.*, 2000). The ITD values also change, since the sound has to bend around the head causing extra propagation delay, but they do not increase as drastically as the ILD values. Note that this effect is not dependent on frequency, and large ILD values are thus generated at low frequencies too. The exaggerated ILD cue compared to the ITD cue thus suggests a source near to the listener.

12.7.2 Accuracy of Distance Perception

A relevant question, then, is: how accurate is the distance perception mechanism? Zahorik *et al.* (2005) analysed 21 studies of perception of distance in different scenarios, and unfortunately found that very different results are obtained with different signals and in different listening conditions. Zahorik *et al.* (2005) concluded that the psychophysical function between the actual distance r and the perceived distance r' fits relatively well with the compressive power function

$$r' = kr^a, \tag{12.3}$$

where k varied between about 0.5 and 2, and a was most typically between 0.3 and 0.8 in the different research results analysed. In practice, the distances are underestimated more or less

severely when the source is farther than about 3 m from the listener, and at distances less than 1 m, the distance of the auditory event overestimates the distance of the sound source.

Zahorik *et al.* (2005) also summarized the directional blur in different studies. In some cases, the blur was only 5–25% of the source distance, while in some other studies the error range was up to 60% to the distance to the source. Higher errors are obtained especially with distant sources.

The hearing mechanisms which are used in distance perception are not known in general. For example, it is not known precisely how the perceived distance is extracted from the effect of the room response in the ear canal signals. A research topic for auditory modelling would thus be to form signal-driven auditory models that measure the cues for distance perception and then combine them to form the final perception of distance.

Summary

This chapter introduced spatial hearing and related concepts. When a sound wave arrives at the listener, the head, the torso, and the external ear affect the sound signals entering the ear canals. These effects are included in head-related transfer functions measured from the ear canals. Based on this information, the auditory system analyses the signals and derives localization cues. The cues, such as interaural differences and the monaural spectrum, are used in associating direction, distance, and other spatial attributes to auditory events and to the internal representation of the acoustic environment. This enables the localization of sound sources with remarkable accuracy even in reverberant spaces.

Listeners are able, at least partially, to compenstate for the frequency response of the listening room; that is, our hearing tries to estimate the signal emitted by the sources without the effect of the room. Humans also have the ability to listen selectively to a certain direction in binaural listening. In multi-source situations, the sources are audible with a 0–8 dB lower level compared to monaural listening.

Further Reading

The book '*Spatial Hearing*' by Blauert (1996) is a broad overview of the field of spatial hearing. The reader might also find the books by Begault (1994), Gilkey and Anderson (1997), and Xie (2013) interesting, especially for HRTF technologies and for applications with virtual displays. Neural processing in humans as well as in some animals is described by Yost and Gourevitch (1987).

References

Alais, D. and Burr, D. (2004) The ventriloquist effect results from near-optimal bimodal integration. *Current Biol.*, **14**(3), 257–262.

Algazi, V.R., Duda, R.O., Thompson, D.M., and Avendano, C. (2001) The CIPIC HRTF database *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop*, pp. 99–102 IEEE.

Begault, D.R. (1994) *3-D Sound for Virtual Reality and Multimedia*. AP professional.

Bilsen, F.A. (1977) Pitch of noise signals: Evidence for a "central spectrum". *J. Acoust. Soc. Am.*, **61**(1), 150–161.

Blauert, J. (1997) *Spatial Hearing – Psychophysics of Human Sound Localization*. MIT Press.

Blodgett, H.G., Wilbanks, W.A., and Jeffress, L.A. (1956) Effect of large interaural time differences upon the judgements of sidedness. *J. Acoust. Soc. Am.*, **28**(4), 639–643.

Borss, C. and Martin, R. (2009) An improved parametric model for perception-based design of virtual acoustics. *35th Int. Audio Eng. Soc. Conf.: Audio for Games*. AES

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.

Brüggen, M. (2001) Coloration and binaural decoloration in natural environments. *Acta Acustica United with Acustica*, **87**(3), 400–406.

Culling, J.F., Colburn, H.S., and Spurchise, M. (2001) Interaural correlation sensitivity. *J. Acoust. Soc. Am.*, **110**(2), 1020–1029.

Damaske, P. and Wagener, B. (1969) Directional hearing tests by the aid of a dummy head. *Acta Acustica United with Acustica*, **21**(1), 30–35.

Duda, R.O. and Martens, W.L. (1998) Range dependence of the response of a spherical head model. *J. Acoust. Soc. Am.* **104**, 3048–3058.

Faller, C. and Merimaa, J. (2004) Source localization in complex listening situations: Selection of binaural cues based on interaural coherence. *J. Acoust. Soc. Am.*, **116**(5), 3075–3089.

Gilkey, R.H. and Anderson, T.R. (eds)(1997) *Binaural and Spatial Hearing in Real and Virtual Environments.* Lawrence Erlbaum Associates.

Gómez Bolaños, J. and Pulkki, V. (2012) HRIR database with measured actual source direction data. *Audio Eng. Soc. Convention 133* AES.

Goupell, M.J. and Hartmann, W.M. (2007) Interaural fluctuations and the detection of interaural incoherence. III. Narrowband experiments and binaural models. *J. Acoust. Soc. Am.*, **122**, 1029–1045.

Grantham, D.W. (1995) Spatial hearing and related phenomena. *Hearing*, **6**, 297–346.

Hafter, E. and De Maio, J. (1975) Difference thresholds for interaural delay. *J. Acoust. Soc. Am.*, **57**(1), 181–187.

Hofman, P.M., Van Riswick, J.G., and Van Opstal, A.J. (1998) Relearning sound localization with new ears. *Nature Neurosci.*, **1**(5), 417–421.

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

Middlebrooks, J.C. (1992) Narrow-band sound localization related to external ear acoustics. *J. Acoust. Soc. Am.*, **92**, 2607–2624.

Mills, A. (1958) On the minimum audible angle. *J. Acoust. Soc. Am.*, **30**(4), 237–246.

Møller, H. (1992) Fundamentals of binaural technology. *Appl. Acoust.*, **36**(3), 171–218.

Nielsen, S.H. (1993) Auditory distance perception in different rooms. *J. Audio Eng. Soc.*, **41**(10), 755–770.

Pellegrini, R. (2002) Perception-based design of virtual rooms for sound reproduction *22nd Int. Audio Eng. Soc. Conf.: Virtual, Synthetic, and Entertainment Audio*. AES

Santala, O. and Pulkki, V. (2011) Directional perception of distributed sound sources. *J. Acoust. Soc. Am.*, **129**, 1522.

Sayers, B.M. (1964) Acoustic-image lateralization judgments with binaural tones. *J. Acoust. Soc. Am.*, **36**(5), 923–926.

Shinn-Cunningham, B. (2000) Learning reverberation: Considerations for spatial auditory displays. *Proc. Int. Conf. on Auditory Display*, pp. 126–134 ICAD.

Shinn-Cunningham, B.G., Santarelli, S., and Kopco, N. (2000) Tori of confusion: Binaural localization cues for sources within reach of a listener. *J. Acoust. Soc. Am.*, **107**(3), 1627–1636.

Stevens, S.S. and Newman, E.B. (1936) The localization of actual sources of sound. *Ame. J. Psychol.*, **48**(2), 297–306.

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

Toole, F. and Sayers, B.M. (1965) Lateralization judgments and the nature of binaural acoustic images. *J. Acoust. Soc. Am.*, **37**(2), 319–324.

von Békésy, G. (1949) The moon illusion and similar auditory phenomena. *Ame. J. Psychol.*, **62**(4), 540–552.

Weiping, T., Ruimin, H., Heng, W., and Wenqin, C. (2010) Measurement and analysis of just noticeable difference of interaural level difference cue. *Int. Conf. Multimedia Technology*, pp. 1–3 IEEE.

Xie, B. (2013) *Head-Related Transfer Function and Virtual Auditory Display*, volume 2. J. Ross Publishing.

Yost, W.Y. and Gourevitch, G. (eds)(1987) *Directional Hearing*. Springer.

Zahorik, P., Brungart, D.S., and Bronkhorst, A.W. (2005) Auditory distance perception in humans: A summary of past and present research. *Acta Acustica United with Acustica*, **91**(3), 409–420.