14

Sound Reproduction

Sound reproduction denotes the process of recording, processing, storing, and recreating sound, such as speech, music, or other sounds. When recording an acoustic scenario, one or more microphones are used to capture sound in single or multiple positions for a recording device. When recording electronical or digital sound sources, microphones are not necessarily required, since the recording devices can directly store the electrical or digital signals. The signals may be processed and stored, and finally made audible to a human listener with loudspeakers or headphones. Note that during this process many choices must be made in order to capture, process, and play back the sound. However, the unifying factor is the common endpoint of the chain, the human listener. This is a distinctive property of the field of communication acoustics compared to some other fields of acoustics. In communication acoustics, the sound is the desired signal, which brings some information or added value to the listener.

Historically, reproduction of sound has come a long way from the first phonograph built by Thomas Edison in 1877. A large variety of sound reproduction applications is currently in use, and they differ both in implementation and the purpose for which they were developed. We begin this chapter by discussing the needs and challenges faced in sound reproduction and continue to present various solutions and technologies for this purpose.

14.1 Need for Sound Reproduction

A wide variety of applications in which sound needs to be reproduced such as:

- *Public address* – the sound amplification systems that make speech or sound audible to a large audience indoors or outdoors.
- *Full-duplex speech communication* over technical channels such as in telephone or teleconference systems.
- Audio content production for the *music and cinema industries.*
- *Broadcasting* of sound in radio or of audiovisual content in TV.

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

- *Computer games and virtual reality*, where sound is captured to be reproduced in association with virtual objects.
- *Accurate reproduction of sound*, where as authentic a replica of the perception of the original sound event as possible is the goal for technical or scientific purposes.
- *Enhancement of acoustics and active noise cancellation*. Sound is captured with a microphone and immediately reproduced after possible processing, resulting in a change in the properties of the sound field. For example, the reverberation properties of a room can be changed, or the noise level can be attenuated.
- In *aided hearing*, devices and processes are used to make sound more audible for the hearing-impaired subject, usually with the aim of making speech more intelligible.

The applications are very different, and the required technical specifications for acceptable functioning may also be very different. Starting from the most basic cases, in certain environmental monitoring systems where the goal is simply to be aware of some sound events, very large deviations in the magnitude and the phase spectrum of the reproduced sound may be allowed. When the requirements are more demanding, the acoustic attributes of the original sound event have to be transmitted with higher fidelity to the listener. For example, with minimum requirements in speech communication, speech is barely understandable without the delivery of prosodic or personal features. In general, the better the original sound signals are preserved in sound reproduction, the higher the quality of reproduction obtained.

In some special cases, full authenticity is sought in reproduction of sound. In communication acoustics, authenticity is determined with human perception as the reference; for authentic reproduction, the auditory perception of a sound scenario in the reproduced conditions should be identical to those in the original conditions, as illustrated in Figure 14.1. Some example cases where authentic reproduction is essential are the reproduction of natural sound scenarios for hearing aid testing and audiology, different training tasks, telepresence applications, and the evaluation of room acoustic parameters of different venues.

The terms *immersion* and *immersive* are often used in connection with sound reproduction, especially in the context of computer games (Björk and Holopainen, 2005), to mean that the

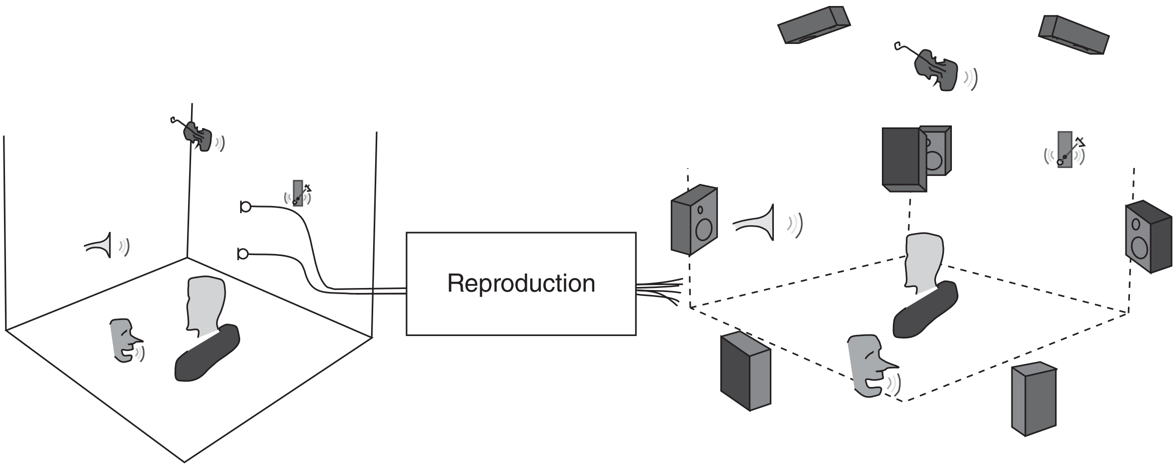

Figure 14.1 The sound reproduction set-up targeting faithful reproduction of an acoustic scenario. Listeners on the left and right ideally should perceive the auditory events identically, that is, the pitch, loudness, duration, timbre, and spatial characteristics of the auditory events should match.

sound reproduced is perceived as convincing. The subject feels that he or she is really ‘in’ the reproduced sound scenario, and the scenario is perceived as if the reproduced sources were real (Kyriakakis, 1998).

14.2 Audio Content Production

The term *audio content* refers here to sound signals produced that have meaning or value to a listener. For example, music recorded in the studio is audio content that can be delivered to a listener. The term *audio engineering* refers to the production of audio content. An *audio engineer* concerns himself or herself with the recording, manipulation, mixing, mastering, and reproduction of sound. Audio engineers creatively use technologies to produce sound for radio, television, film, public address, electronic publishing, and computer games.

Some important terms used in audio engineering that must be understood are:

- *Recording* is the process of capturing sound in a real acoustic scenario, or capturing the output signals of electrical or digital sources. Capturing of sound from an acoustic scenario can be conducted using one or more microphones, close to or far from the sound source, indoors or outdoors, and storing it in a recording device. Typically, each microphone signal is stored on one channel, or *track*, of the device. The capturing of digital and electrical signals is trivial with such devices, as they have electrical and/or digital inputs.
- *Mixing* is the process of adding different recorded tracks together after they are amplified, and possibly processed with systems that modify audio signals (audio effects). *Audio effects* are, for example, equalization, dynamic range control, panning, and reverberation. The following sections describe these effects in more detail. Mixing is done either on a *mixing console*, which is a dedicated device for this process, or using dedicated computer software. The outcome of mixing is a single- or multi-channel audio track that is meant to be played over a *loudspeaker set-up*, where each loudspeaker reproduces one channel.
- *Mastering* is the process of preparing and transferring the mixed audio track to its final form, ready to be copied to media, such as a CD, DVD, or the Internet, or for broadcasting. The sound is often processed further during mastering, where typically the sound signals are equalized and some dynamic range control is applied.
- *Live sound* is the on-line mixing and mastering of the audio signals by audio engineers during live concerts or live broadcasts for the public-address loudspeakers, for the signal stream for the purpose of broadcasting, or for both. The audio systems are typically set up earlier, and the desired settings for devices are found during a *sound check*.
- *Studio* – a facility that generally consists of at least two rooms: the studio(s) or *live room(s)*, where the music or other sound is generated, and the *control room(s)*, to where the sound signals from the microphones in the live rooms are routed, stored, and later mixed. A room may also be dedicated to the mastering process.
- *Audio format* defines the number of tracks, the loudspeaker set-up, and the encoding/decoding method used to record the audio content. In most cases, each track is meant to be listened to using a dedicated loudspeaker, although some exceptions exist.

Audio content production seldom strives to relay faithfully the listening experience of a real acoustic scenario, such as a musical concert. Even the recordings from live concerts are processed to produce a ‘good-sounding’ and plausible result, and typically an authentic reproduction of a live event is not the goal. By *plausible* we mean that the perceived features of the reproduced audio are believable and correspond well with listeners' expectations in the given context.

Audio engineering is a form of art, where an audible piece of art – music, speech, or other soundscape – is turned into audio tracks, ready for later listening. For authentic reproduction of audio content, the final listening experience created by the audio engineer in the mastering (or mixing) studio should be conveyed to listening consumers.

14.3 Listening Set-Ups

Sound can be made audible over different loudspeaker set-ups, ranging from a monophonic set-up to multi-channel systems and headphones. The best-known systems will be discussed below. The acoustic effect of the listening room also affects the perception of the reproduced sound. The loudspeaker set-up may also be accompanied by a visual display and/or vibroacoustic transducers, which add different cross-modal effects to the process.

14.3.1 Loudspeaker Set-ups

A *loudspeaker set-up*, or *loudspeaker layout*, defines the number of loudspeakers and their directions with respect to the *best listening position*. The term *listening area* refers to the area where the system is listened to. The distances from the best listening position to the loudspeakers and the responses of each loudspeaker are assumed to be identical. If they are not identical, some delaying and gaining may be applied to loudspeaker channels to compensate for the differences.

The most common loudspeaker set-up is the two-channel stereophonic set-up. Its use became widespread after the development of the single-groove 45°/45° two-channel record in the late 1950s. Two loudspeakers are positioned in front of the listener 60° apart, as illustrated in Figure 14.2a. The set-up enables the positioning of virtual sources between the loudspeakers, and it also makes the timbral quality of reproduction better when compared to monophonic reproduction, as discussed later in Section 14.4.1.

The motivation for using more than two loudspeakers in the reproduction is the potentially better spatial quality in a larger listening area. Different multi-channel loudspeaker set-ups have been specified in the history of multi-channel audio (Davis, 2003; Steinke, 1996; Torick, 1998). In the 1970s, the quadraphonic set-up was proposed, in which four loudspeakers are positioned evenly around the listener at azimuth angles ±45° and ±135°. This layout was never successful because of problems related to content delivery techniques at that time, and because the layout itself has too few loudspeakers to provide good spatial quality in all directions around the listener (Rumsey, 2001).

A sound reproduction system was developed for cinema whereby the front image stability of the standard stereophonic set-up was enhanced by an extra centre channel and two surround channels were added to create atmospheric effects and room perception. This surround sound system for cinemas was first used in 1976 (Davis, 2003), and the ITU made a recommendation for the layout in 1992 (BS.775-2, 2006). The late 1990s saw households also acquiring this 5.1 surround system, where the figure before the dot stands for the number of loudspeakers and the figure after the dot is the number of low-frequency channels. In the ITU recommendation, the three frontal loudspeakers are in the directions 0° and ±30°, and the two surround channels

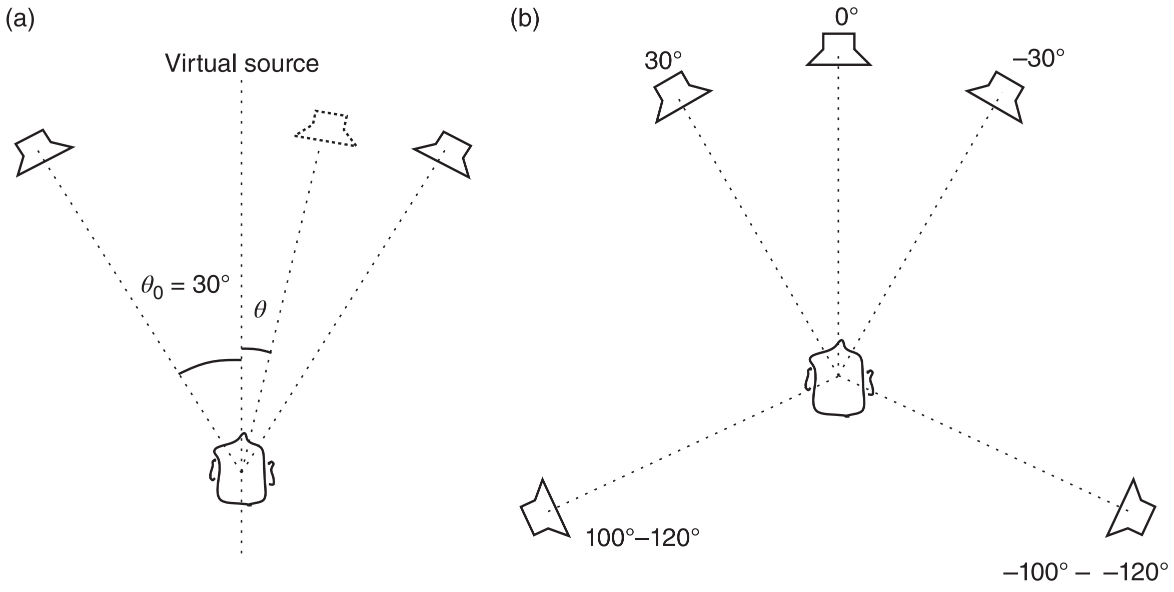

Figure 14.2 (a) The standard stereophonic listening configuration. (b) The 5-channel surround loudspeaker set-up based on the ITU recommendation BS775-2.

in directions ±110 ± 10°, as shown in Figure 14.2b. The system has been criticized for not being able to deliver good directional quality anywhere else other than in the front (Rumsey, 2001). So other layouts with 6–12 loudspeakers have been proposed to enhance the directional quality in other directions as well.

A factor limiting the use of loudspeakers is their physical size. Optimally, the loudspeakers should be relatively large in order to be able to reproduce low frequencies. The smaller the loudspeaker, the higher is the lowest frequency it can reproduce. Unfortunately, larger size also means increased costs and more complicated installation. The idea of using a subwoofer(s) is to reproduce the low frequencies of the stereophonic or multi-channel audio track using one or many loudspeakers dedicated for reproducing frequencies typically under 80–200 Hz (Borenius, 1985; Welti, 2004). Broadband loudspeakers can then be designed to be relatively small without the need to reproduce low frequencies. The motivation for this design comes from psychoacoustics. Our directional hearing at low frequencies is rather poor, and thus not having broadband loudspeakers reproducing low frequencies is assumed not to severely impair the overall quality of reproduction.

All the loudspeaker layouts described above have loudspeakers only in the horizontal plane. However, there are systems, for use in theaters and virtual environments, in which loudspeakers are placed above and/or below the listener too, thus enhancing the perceived realism, especially in situations where the 3D position of a virtual source is important (Silzle *et al.*, 2011). Typical examples of such situations are virtual sources for flying vehicles or the sound of raindrops on a roof. Such 3D set-ups have been proposed for use in domestic listening too, and are currently being standardized (ISO/IEC 23008-1, 2014). For example, Japan Broadcasting Corporation has proposed a 22.2 loudspeaker set-up (Hamasaki *et al.*, 2005), which has 22 loudspeakers in planes at three heights and two subwoofers, or a common set-up has four elevated loudspeakers added to the 5.1 set-up. Note that the positioning of loudspeakers in domestic or other real listening set-ups may be very different from the theoretical specifications of the layouts. For example, the loudspeakers of a stereophonic set-up may, in practice, be in any direction and at arbitrary distances from the listener. In addition, in cars the loudspeaker set-ups are different for different car models, and none of the passengers are situated equidistantly from the loudspeakers. The audio content thus has to be created in such a way that such deviations from the theoretical set-ups do not affect the listening experience too much.

Some audio formats meant primarily for cinema also allow the use of different multi-channel systems to reproduce the audio content. The audio is sent with some metadata, which then defines how the sound is rendered to the loudspeakers. The corresponding methods are called *object-based audio* techniques. Such systems, like those covered by Robinson *et al.* (2012) and Lemieux *et al.* (2013), are called *loudspeaker-set-up agnostics*. The first movies with sound tracks in such formats were released in 2012.

14.3.2 Listening Room Acoustics

The acoustics of listening rooms vary significantly. Car audio systems and headphones have very short or non-existent reverberation. Small rooms and big rooms produce very different room responses. These differences make the goal of providing an identical listening experience in every listening environment appear impossible to achieve. However, despite the great variations in listening room acoustics, a certain piece of audio ‘sounds’ very similar in many different listening conditions. Listeners are typically able to identify the sound sources of a musical piece, recognize, say, the vocalist, the lyrics, and the music itself correctly, although the signals in their ear canals may have very different spectral content in different rooms (Toole, 2012). The ability of human hearing to adapt to diverse listening room acoustics is remarkable, as discussed briefly in Sections 11.5.1 and 12.6.

As a matter of fact, mixing and mastering studios around the world have different acoustics and use different audio devices. In principle, the audio content should produce a similar perception of sound to how the audio engineer perceived it in the final stage of the audio content production, as reasoned at the end of Section 14.1. But since each and every studio has a different listening set-up, there is no single answer to the question of what kind of acoustic properties domestic listening rooms should have ideally. However, the fact that humans are able, at least partly, to eliminate the effect of room acoustics in timbre perception mitigates the issue.

The acoustics of the listening room may have larger implications on certain perceptual attributes. For example, the perception of spatial characteristics of sound is largely affected by the listening room. The perceived directions of sources may be smeared if the room has strong early reflections, and the perception of reverberation in recorded sound is generally impossible if the listening room has stronger reverberation in itself. In addition, certain non-linear distortions, such as pre-delays and smearing of transients that may occur in lossy perceptual coding, may be more audible in less reverberant rooms. Our current knowledge of the effect of the listening room and loudspeaker characteristics on listening experience is summarized by Toole (2012) and Bech and Zacharov (2006).

A number of specifications for listening room acoustics and loudspeaker set-ups therein have been proposed. The main purpose for such specifications is to make the results of psychoacoustic listening tests conducted in rooms comparable between different academic and industrial sites. The specifications are relatively detailed. For example, ITU-R BS.1116-1 (1997) gives the details for the geometry of the room, the reverberation time, the amount of allowed background noise, the loudspeaker set-up geometry, listening positions, and audio system performance. Most of the values are given with tolerance rates, like how much the reverberation time can deviate from the optimal value depending on frequency. The resulting rooms resemble living rooms with relatively short reverberation time T_{60} (see Section 2.4.2 for a definition of T_{60}). T_{60} is also required to be almost constant over a large frequency range.

14.3.3 Audiovisual Systems

Sound is often reproduced in the presence of a moving picture, as in TV or cinema. Historically, starting from the 1890s, the first three decades of the cinema industry were the era of 'silent movies'. The technological advances in sound reproduction in the late 1920s made it possible to synchronize monophonic sound with the moving picture, soon replacing the silent movies with 'talking movies'. Cinema sound reproduction has progressed through two-channel stereo in the 1950s to the delivery of multiple discrete loudspeaker playback channels in the 1990s, and in the 2010s to loudspeaker-set-up-agnostic object-based audio formats. New developments of audiovisual systems have often been taken into use first in the cinema, and later they have been introduced for domestic use too.

The effect of the degradation of different features of sound on the quality of audio and video has been studied a lot, and an overview of these studies is given by Kohlrausch and van de Par (2005). The studies show that the effect of degradation on the quality of either audio or video depends on the content, implying that, in some cases, degradation in audio is more easily noticed than in video and vice versa (Hollier *et al.*, 1999). It is quite natural that degradation is more audible in the modality on which the subject is focusing (Rimell and Owen, 2000). Cross-modal effects also exist. Joly *et al.* (2001) showed that better audio quality can make video degradation less annoying, but good video quality was not found to improve the perceived audio quality.

A common problem found in audiovisual reproduction is the lack of synchronicity between audio and video. The lack of synchronicity is often caused because the audio and video signal routes produce different delays. The detection of asynchronicity is not symmetric across the modalities. Audio that is presented too early is noticed when the time shift is shorter than when it is presented after the corresponding video. Different studies propose different numbers for the threshold of detection of the shift: an audio lead of 25–75 ms and lag of 40–90 ms are said to be detected (Levitin *et al.*, 2000). This can be understood from the observation that in nature, for larger distances between the subject and the source, sound arrives at the ears considerably later than light at eyes from the source. Thus, humans are quite accustomed to sound stimuli lagging visual stimuli. The ITU has published a recommendation concerning synchronization thresholds for audio and video components in television signals (ITU-T, 1990). The maximum tolerated lead of audio in the recommendation is 20 ms, and correspondingly the maximum tolerated lag is 40 ms.

The presentation of both audio and video to the subject creates many types of interactions between modalities. For example, the sound of a red car is perceived to be louder than a blue or green one, with a difference corresponding to 1–3 dB of SPL (Menzel *et al.*, 2008). The audio track also influences eye movements and the direction of gaze, and sound has been shown to strengthen the salience of corresponding visual events (Coutrot *et al.*, 2012).

The perception of audio and video alone often differ in their spatial characteristics. Auditory objects may be localized to a position different from their visual counterparts, and the space perceived visually may be larger or smaller than that perceived by hearing. When the audio and video are reproduced simultaneously, the perceptual mechanisms try to resolve such conflicts. Ventriloquism exploits how our brain resolves these conflicts, as already discussed in Section 12.3.5 on page 234, where it was shown that correlated sound and video objects are perceived to be in the most probable direction, based on the cues available.

14.3.4 Auditory-Tactile Systems

Sound can also be perceived through the sense of touch. The human skin has a large number of mechanoreceptors, which are sensitive to vibrations. The vibration frequencies perceived by humans range from a few hertz to a few hundred hertz. The vibrations perceived by humans exist in diverse situations, from inside vehicles to live concerts. Some musical instruments also make the structures in the listening room vibrate. The airborne sound and structural vibrations are then transferred to the listener and may produce tactile perceptions. Examples include PA systems in loud rock concerts, timpani, and church organs. However, listeners may not consciously notice the vibrations, as they are typically in synchronicity with the music or other sounds, so that different percepts fuse, causing a holistic perception of events.

The degree of accuracy in the perception of the frequency spectrum of vibrations is nothing near that of the auditory system. However, humans are able to perceive the presence of vibration, and its frequency is analysed with very low selectivity. The JND of the level of stimulus is thought to be of the order of 1 dB, which is similar to that for auditory perception. In frequency-matching tasks, subjects identify relatively accurately or misjudge to be an octave lower a tactile sinusoidal stimulation between 60 and 180 Hz (Altinsoy and Merchel, 2010).

The interaction between presented tactile and auditory stimuli is of interest in this book. The loudness of a sound presented via headphones is perceived to be higher in the presence of a simultaneous tactile vibration produced with a whole-body shaker. The change in loudness corresponds to a change in level of about 1 dB (Merchel *et al.*, 2009). In bass reproduction, the preferred level of acoustic sound is lower if a shaker is also used to reproduce low frequencies (Simon *et al.*, 2009). Interestingly, this auditory–tactile loudness interaction does not depend significantly on the power of the vibration (Merchel *et al.*, 2009).

The JND of detection of haptic–audio asynchronicity has been found to be 24 ms with impulsive stimuli (Adelstein *et al.*, 2003), which is a value similar to that obtained with tests on the detection of audio–visual asynchronicity. However, in many cases, the auditory–visual asynchronicity threshold values are higher than auditory–tactile asynchronicity threshold values. Therefore, auditory–tactile asynchronicity may be even more critical than auditory–visual asynchronicity. In particular, musicians have lower auditory–tactile asynchronicity thresholds, about 10 ms, than the general population, possibly because of their training.

There are some interesting applications designed for auditory-tactile systems. The perceived quality of music reproduction has been found to be higher when vibrations are reproduced to the listener through a chair (Merchel and Altinsoy, 2013). A haptic device may also be used to provide the user with feedback. For example, mixing desks typically have a large number of input channels, each having a fader to control the level of the associated instrument. If the sound of the instrument is used to vibrate the faders, the audio engineer can recognize the instrument in each channel simply by touching the fader. This enables heads-up mixing, or mixing in low-light conditions (Merchel *et al.*, 2010).

14.4 Recording Techniques

The term *recording technique* is used here rather loosely to refer to the number of microphones used and how they are positioned in relation to each other, to the sound source(s), and to the recording room. A technique may also specify how the recorded signals are processed and how they are routed to the loudspeaker(s). The most common recording techniques are described below. The largest differences between them are in how they reproduce the spatial characteristics of sound. The perception of a reproduced spatial sound scene depends significantly on how the microphones are positioned and how their signals are processed and routed to loudspeakers.

14.4.1 Monophonic Techniques

The simplest recording technique is monophonic recording, where one or more microphones are used near to or far from the sources. The single recorded track is reproduced over a loudspeaker, or the recorded signals are mixed together and reproduced via a single loudspeaker. The reproduction is called monophonic even if this single signal is applied to a larger number of loudspeakers, such as in public address or general paging systems. The non-spatial characteristics of the sound can be captured using the monophonic recording technique, which is often good enough for many applications, such as speech communication.

A major disadvantage of monophonic reproduction is that the colouration caused by the recording room is exaggerated in the listening when compared to binaural listening of the recording venue. The reason for this is that the reproduced single sound signal gets filtered by the room, manifesting as a complex structure in the frequency response, as illustrated in Figure 2.23 on page 39. The sound reaching the two ears of a listener through the room has *different* magnitude and phase spectra, which results in binaural decolouration, as discussed in Section 12.6.2. The spectrum emanated by the single loudspeaker during monophonic reproduction is already coloured by the recording room response, but the binaural decolouration mechanisms can only try to compensate for the *listening room* acoustics, not the *recording room* acoustics. This emphasizes the acoustic effect of the recording room. When the recording is made with at least two microphones and reproduced over at least two loudspeakers, the recording room effect is different in the different loudspeaker signals, which enables at least partial decolouration of the recording room effect.

However, monophonic sound reproduction is evidently the most used sound reproduction method, since it is used in telephony. When the microphone is at a distance of a few centimetres from the source, the mouth of the person speaking, the acoustics of the recording room do not have much of an effect on the result, and the timbral quality of the reproduction is good. The disadvantage discussed in the previous paragraph is thus valid only when the captured effect of the room is significant, or, in technical terms, the level of the direct sound is comparable to or lower than that of the reverberant field.

14.4.2 Spot Microphone Technique

The *spot microphone technique* (also called close miking) records a number of channels, and the final audio content is mixed by an audio engineer. The technique is commonly used to capture a concert or studio session where many instruments are played. A microphone, called a *spot microphone*, is placed near each source to be captured (Rumsey, 2001). The microphone signals are often required to be as independent as possible, each signal containing sound from one source only.

The positioning of the microphones is critical, since the frequency-dependent directional patterns of many instruments have a major effect on the magnitude spectrum of the sound radiated in different directions (Pätynen and Lokki, 2010). Such recording is often complemented with distant microphones, which are used to record the sound from the sources through the response of the room. Such microphones are often called *ambience* or *ambient* microphones. The spot microphone signals together with the ambience microphone signals are then used to create the final audio content through the mixing process.

14.4.3 Coincident Microphone Techniques for Two-Channel Stereophony

Coincident microphone techniques for two-channel stereophonic systems are often referred to as *XY techniques*. Two directional microphones are placed with their diaphragms as close to each other as possible, but facing different directions. Typically, two cardioid or hypercardioid microphones are used, and the angle between the directions of the microphones is −60° − 120°, although angles up to 180° may also be used (Streicher and Dooley, 1985). Each of the recorded signals is applied to the corresponding loudspeaker of the stereophonic set-up. The recorded sources perceptually appear relatively point-like in the reproduced sound (Lipshitz, 1986) when listened to in the best listening position. When the listener moves away from the best listening position by a few tens of centimetres, the perceived sources migrate to the nearest loudspeaker.

However, these techniques have been characterized to lack a 'sense of space' (Streicher and Dooley, 1985). One reason for the weak perception of space is that the microphones are directional and are directed typically towards the instruments. The effect of the room, in turn, arrives evenly from all directions, and thus the directional pattern effectively reduces the level of reverberation in the recording.

The first XY techniques were presented by Blumlein (1931) in the 1930s. A particular XY technique is still known as the *Blumlein pair*, where the positive lobes of the two dipole microphones are in the directions 45° and −45° azimuth. The sound source(s) to be recorded is situated between the directions of the positive lobes. The technique reproduces the sound energy equally in all directions and does not suffer from attenuated capture of the reverberant field. However, the sound arriving from the sides produces a 180° phase reversal between the microphone signals. The sound is thus reproduced in loudspeakers out of phase, which may produce colouration and directional artefacts.

A microphone technique widely known as *MS stereo* was also presented by Blumlein (1931), where a directional microphone (M) is directed towards the sound source(s) and a figure-of-eight microphone (S) to the side. Then, a weighted sum of the microphone signals is made for each loudspeaker, thus creating two virtual microphone signals. Such a computation is also called *matrixing*. The directional patterns of these virtual microphone signals can be adjusted during mixing by changing the weights in the summing, as explained in more detail in the context of Ambisonics in Section 14.4.6.

14.4.4 Spaced Microphone Techniques for Two-Channel Stereophony

Spaced microphone techniques differ from coincident techniques in that the microphones are typically placed 20 cm to a few metres apart. These techniques are also known as *AB techniques*. Often, two microphones with omnidirectional patterns are used, although directional microphones may also be utilized. The signals from each source arrive at different times at the two microphones, and thus the time difference between the microphone signals depends on the direction of arrival. Also, differences in amplitude might exist if the microphones are far away from each other.

If the recorded signals are impulse-like, virtual sources may be perceived to be point-like and arriving from a direction that depends on the time delays between the microphone signals.

For tonal signals, such as harmonic complexes, the virtual sources may be localized inconsistently; the localization varies with frequency, and a virtual source may be perceived to be wide. However, sound recorded with spaced microphone techniques is often considered to be more 'ambient', 'airy', or 'warm' than sound recorded with coincident microphone techniques (Lipshitz, 1986). A reason for the more pronounced room response in AB techniques than in XY techniques is the omnidirectional response, which does not attenuate the diffuse field. Another effect that emphasizes the room response is that, due to the considerable distance between the microphones, the captured signals are incoherent within a 1-ms window for most directions of arrival. When such incoherent signals are applied to loudspeakers, the interaural cues of the listener suggest directions exceeding the directions between the loudspeakers in the stereophonic set-up, which may lead to higher detectability of the reverberant sound (Pulkki, 2002).

There also exist methods, such as the *ORTF* technique, that can be classified as falling this technique between the coincident and spaced microphone techniques. In it, two cardioid microphones are positioned 17 cm apart with an angle of 110°. The captured signals may thus differ both in time and in amplitude. At low frequencies, the system corresponds to the XY cardioid technique, since the distance between the microphones is small compared to the wavelength. At high frequencies, the microphone signals have considerable phase differences. The perceptual attributes of the reproduced virtual sources and reverberation are reported to be somewhere between the attributes of the coincident and the spaced techniques (Lipshitz, 1986).

14.4.5 Spaced Microphone Techniques for Multi-Channel Loudspeaker Systems

When recording a live acoustic event for reproduction over a multi-channel loudspeaker set-up, typically a number of microphones are positioned in a spaced configuration at the recording venue. The number of microphones is at least the same as the number of loudspeakers, and they are separated by distances ranging from 10 cm to several metres. Often, the microphones face approximately towards the corresponding loudspeaker directions (Rumsey, 2001). Microphones with different directional patterns are used in the set-ups. Some proposed arrangements that achieve different properties in reproduction of sound exist, such as the layout shown in Figure 14.3a. Different microphone arrangements can also be used near and far from the sources. Such a layout is shown in Figure 14.3b, where several microphones are located close to the sources and a Hamasaki square of four figure-of-eight microphones is located far from the sources, well outside the critical distance. The recorded signals are then mixed and mastered into the final multi-channel audio format.

There is no single microphone technique that is the best solution in all situations. In practice, audio engineers decide microphone positions separately for every recording venue and for every ensemble to be recorded, tuning the set-up using monitoring loudspeakers in a separate listening room. The resulting surround audio content is, at best, plausible and enjoyable, some kind of spaced microphone set-up being the most often used technique for surround sound recordings.

14.4.6 Coincident Recording for Multi-Channel Set-up with Ambisonics

The *Ambisonics* reproduction technique (Gerzon, 1973) provides a theoretical framework for coincident recording techniques for 2D and 3D multi-channel loudspeaker set-ups. In theory, Ambisonics is a compact-format, loudspeaker-set-up-agnostic, efficient, and comprehensive

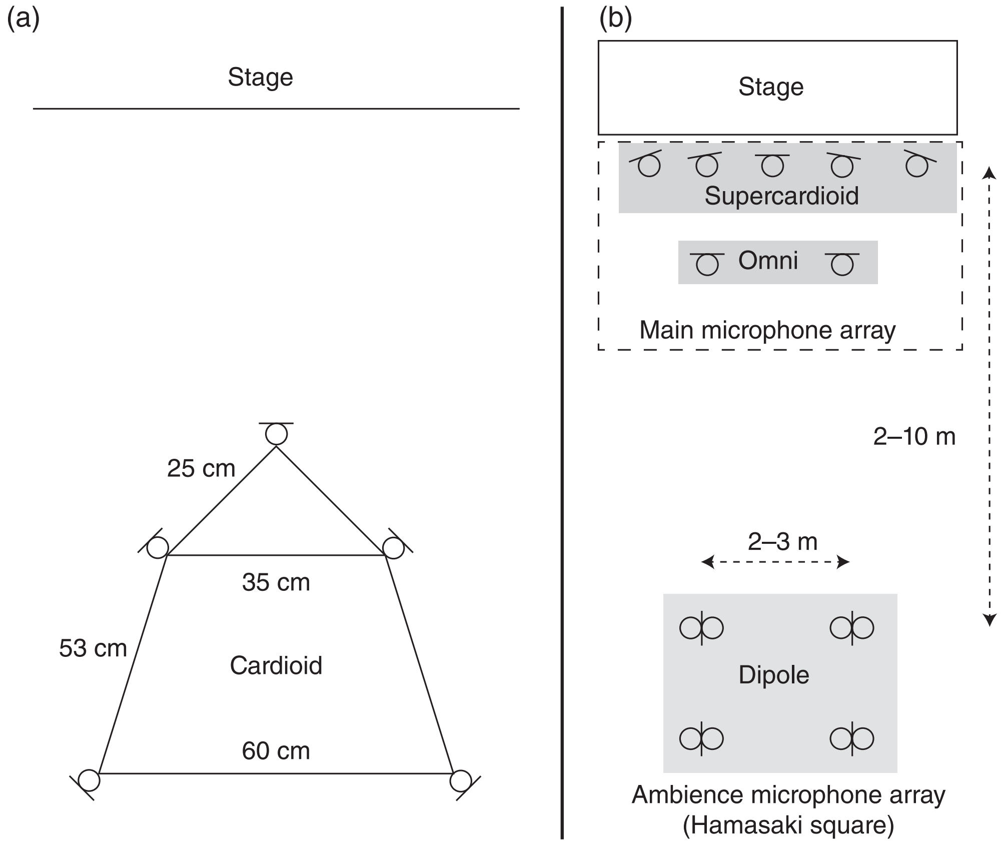

Figure 14.3 Two microphone configurations for recording for 5.1 surround reproduction. (a) The recording of five signals for five loudspeakers using the 'optimal cardioid array', as described by Rumsey (2001). (b) The main microphone array near the sources and an ambience array far from the sources used to record more than five signals, which are then mixed into loudspeaker signals (Hamasaki and Hiyama, 2003).

method for capture, storage, and reproduction of spatial sound. Unfortunately, it has some unsolved technical drawbacks in recording of real acoustic scenarios, which, however, can be at least partly mitigated using non-linear techniques.

All coincident multi-channel microphone techniques produce signals that can be transformed into *B-format* signals. B-format signals have directional patterns that have *spherical harmonics*. The zeroth-order harmonic is the omnidirectional pattern, the three first-order patterns are those of a dipole facing in each of the three directions of the axes of the Cartesian coordinate system, and second order patterns are more complicated, having the form of a quadrupole or a dipole-with-ring form, as shown in Figure 14.4. The directional patterns of signals are additive; when two signals are added, the directional pattern of the combined signal follows the combined directional pattern. For example, combining the omnidirectional and dipole patterns with equal gains produces a signal with a cardioid pattern. When higher-order components are recorded, more complex directional patterns are formed for the combined signals. The spherical harmonics can thus be seen as basis functions for the design of arbitrary patterns.

The most common microphone device for Ambisonics is the first-order, four-capsule *B-format* microphone, producing signals with directional patterns having spherical harmonics up to the first order. The omnidirectional signal is denoted $w(t)$, and the three dipole signals are denoted $x(t)$, $y(t)$, and $z(t)$. The $w(t)$ signal is usually scaled down by a factor of $\sqrt{2}$.

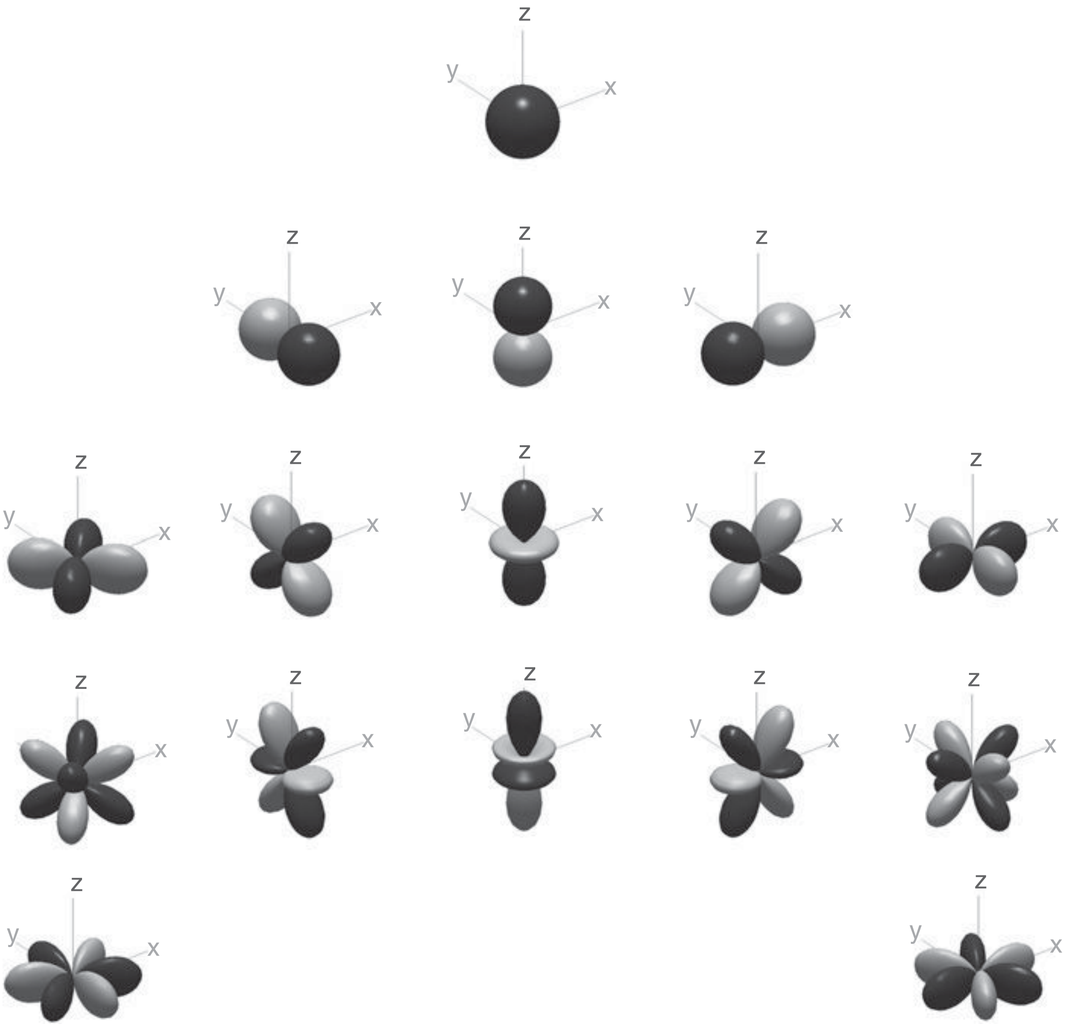

Figure 14.4 From top to bottom: spherical harmonics of order 0, 1, 2, and 3. Third-order sperical harmonics are shown in the two bottom rows. Courtesy of Archontis Politis.

Higher-order microphones with more capsules have also been developed and are commercially available. The higher-order components can then be derived in a specific frequency window. There are some higher-order microphones available which can extract harmonics up to about the fifth order in a certain frequency window. The number of good-quality microphone capsules in such devices is relatively high, something like 32. Outside the frequency window, the microphones suffer from low-frequency noise and deformation of directional patterns at high frequencies (Moreau *et al.*, 2006; Rafaely *et al.*, 2007).

First-order microphones have been available for decades, and we use as an example of how to compute signals for loudspeakers. The channels are matrixed that is, added together with different gains. Thus, each loudspeaker signal can be considered as a virtual microphone signal having first-order directional characteristics. This is expressed as

$$s(t) = (1-\kappa)w(t) + \frac{\kappa}{\sqrt{2}}[\cos(\theta)\cos(\phi)x(t) + \sin(\theta)\cos(\phi)y(t) + \sin(\phi)z(t)], \qquad (14.1)$$

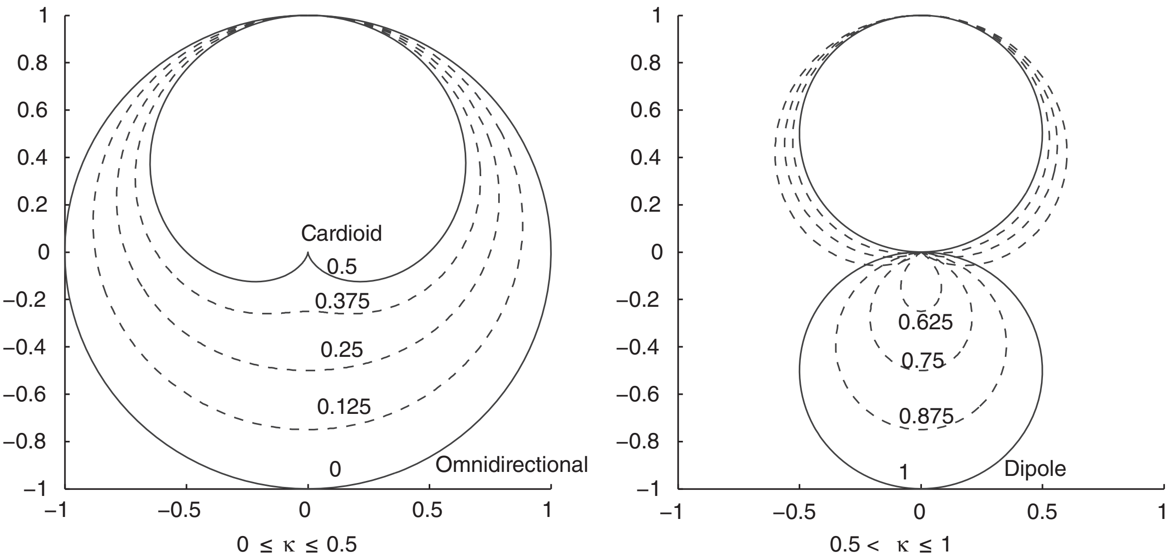

Figure 14.5 The directional patterns of virtual microphones that can be generated from first-order B-format recordings.

where $s(t)$ is the produced virtual microphone signal oriented at azimuth angle θ and elevation ϕ. The parameter $\kappa \in [0, 1]$ defines the directional characteristics of the virtual microphone from omnidirectional to cardioid and dipole, as shown in Figure 14.5.

In principle, first-order Ambisonics could be used for any loudspeaker set-up, but unfortunately it has a very limited range of use. The broad first-order directional patterns make the listening area very small, extending the size of the head of the listener only at frequencies below about 700 Hz (Solvang, 2008). At higher frequencies, the high coherence between the loudspeaker signals leads to undesired effects, such as colouration and loss of spaciousness. The number of loudspeakers in a horizontal set-up should not exceed $2N + 1$, where N is the order of the B-format microphone, to avoid too high a coherence between the loudspeaker signals. Thus, first-order microphones can be used only with three-loudspeaker set-ups, which is far too few to produce the perception of virtual sources between the loudspeakers. This calls for the use of microphone set-ups able to capture signals with higher-order spherical harmonic directional patterns. With higher-order directional signal components, the directional patterns of the loudspeaker signals can be made narrower, which solves these issues.

14.4.7 Non-Linear Time–Frequency-domain Reproduction of Spatial Sound

The spatial resolution of hearing is limited within the auditory frequency bands (Blauert, 1996). In principle, all sound within one critical band can only be perceived as a single source with a broader or narrower extent. The limitations of spatial auditory perception raise the question of whether the spatial accuracy in reproduction of an acoustic wave field can be compromised without a decrease in perceptual quality. When some assumptions on the resolution of human spatial hearing are used to derive reproduction techniques, potentially an enhanced quality of reproduction is obtained.

The audio recording and reproduction technology called *directional audio coding* (DirAC) (Merimaa and Pulkki, 2004; Pulkki, 2007) assumes that the spatial resolution of the auditory system at any one time instant and in one critical band is limited to extracting one cue for direction and another for interaural coherence. It further assumes that if the direction and diffuseness of the sound field are measured and reproduced correctly with a suitable time resolution, a human listener will perceive the directional and coherence cues correctly.

An example of the implementation of DirAC is shown in Figure 14.6. In the analysis, the direction and diffuseness of the sound field are estimated using temporal energy analysis in the auditory frequency bands, as described below. The direction is expressed in azimuth and elevation angles, indicating the most important direction of arrival of sound energy. Diffuseness is a real number between zero and one, which indicates whether a sound field more closely resembles a plane wave or a diffuse field. Virtual microphones are then formed from B-format signals, which are divided into a *diffuse stream* and a *non-diffuse stream* using the diffuseness parameter. The non-diffuse stream is assumed to contain sound that originates primarily from one source, allowing it to be applied in a single direction. This is implemented here by computing amplitude-panning gain factors (see Section 14.5.3) and by gaining the virtual microphone signals with them. The sound is thus effectively reproduced only by loudspeakers near the analysed direction.

The diffuse stream, in turn, is assumed to contain sound originating from reverberation or from multiple concurrent sources from different directions, which should produce low interaural coherence. The virtual microphone signals in the diffuse stream divide the energy around the listener, which fits nicely the assumption of diffuse reverberation. Unfortunately, the virtual microphone signals have too high coherence between each other, which produces similar shortcomings to the first-order Ambisonics. A decorrelation process can be used to reduce the coherence between the channels, so that the phase spectrum of the signals is altered, ideally without affecting the magnitude spectrum, as explained in Section 15.2.8. The shortcoming of decorrelation, time-smearing of transients, is typically not audible, since diffuse sound does not typically contain strong transients.

This implementation of DirAC has been proven to produce better perceptual quality in loudspeaker listening than other available techniques using the same microphone input (Vilkamo *et al.*, 2009). An almost authentic reproduction is obtained if the acoustic scenario fits the DirAC signal model, where sound from one source dominates in one frequency band with potentially some diffuse sound arising from room reverberation. On the other hand, two broadband sources with temporally relatively smooth envelopes that arrive at the microphone from opposite directions cause some quality degradation (Laitinen, 2014). The degradation may be audible as smearing of transients or a perception of added room effect. A number of techniques to avoid such degradation with first-order B-format input have been proposed (Laitinen, 2014; Vilkamo and Pulkki, 2013).

Similar assumptions have been used in the development of other time–frequency-domain methods for spatial sound reproduction (Alexandridis *et al.*, 2013; Berge and Barrett, 2010; Cobos *et al.*, 2010; Tournery *et al.*, 2010). DirAC has also been developed for higher-order B-format microphone input by dividing the sound field into sectors and performing the analysis and synthesis individually for each sector, which makes the acoustic conditions less challenging for the analysis–synthesis methods (Politis *et al.*, n.d.).

A basic method of analysing the sound field to compute the DirAC metadata is described briefly. Let us assume that frequency-domain signals for pressure P and 3D particle velocity

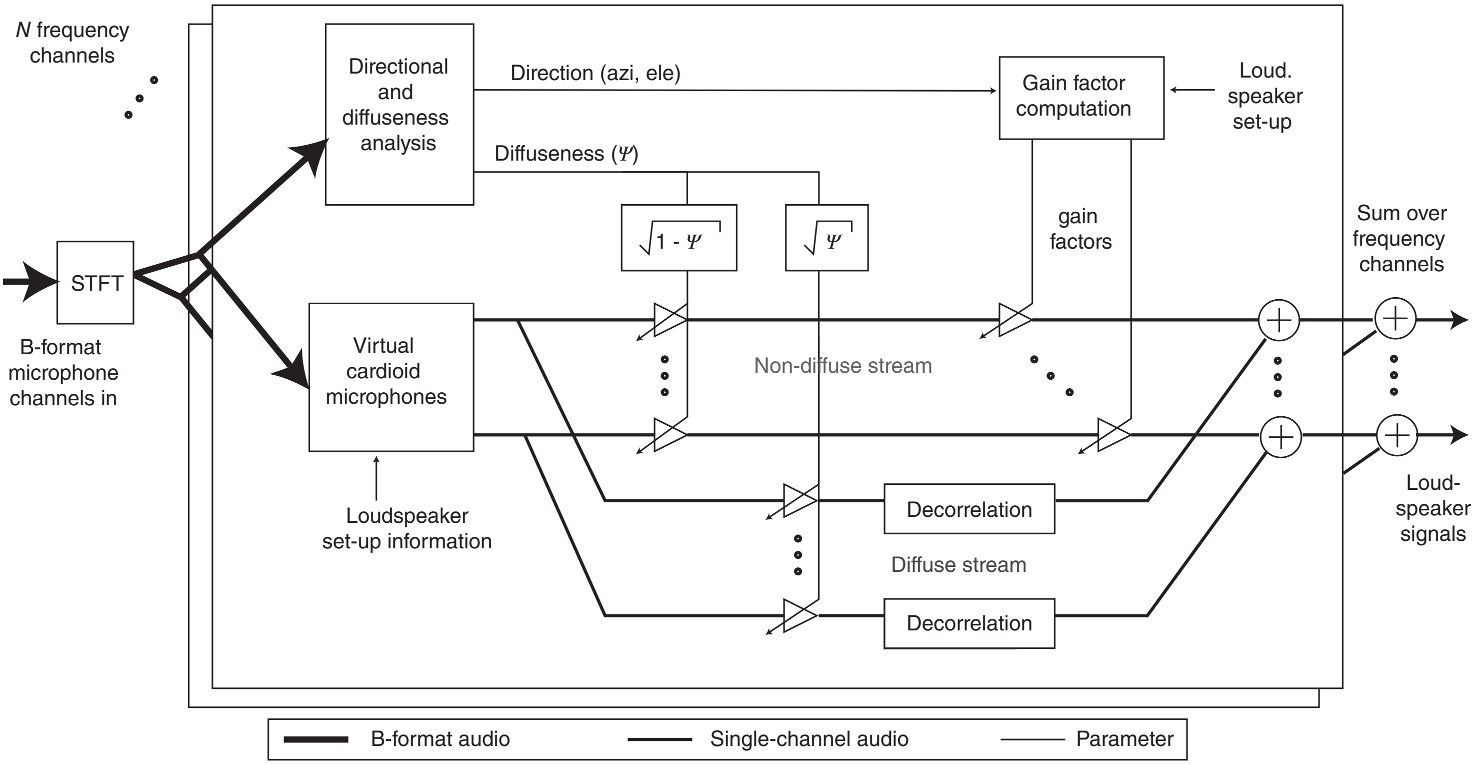

Figure 14.6 Directional audio coding with virtual microphone implementation.

U are available. They can be estimated from B-format signals, but the mathematical derivation of the estimates is beyond the scope of this book. The energy E of the sound field can be computed as

$$E = \frac{\rho_0}{4}||\mathbf{U}||^2 + \frac{1}{4\rho_0 \mathrm{c}^2}|P|^2, \tag{14.2}$$

where ρ_0 is the mean density of air and c is the speed of sound.

The *intensity vector* **I** expresses the net flow of sound energy as a 3D vector, and can be computed as

$$\mathbf{I} = P^*\mathbf{U}, \tag{14.3}$$

where $(\cdot)^*$ denotes complex conjugation. The direction of arrival of sound is defined to be the opposite of that of the intensity vector in each frequency band. The direction is denoted as corresponding angular azimuth and elevation values in the transmitted metadata. The diffuseness of the sound field is computed as

$$\psi = 1 - \frac{|\mathrm{E}\{\mathbf{I}\}|}{\mathrm{c}\mathrm{E}\{E\}}, \tag{14.4}$$

where E is the expectation operator. Typically, the expectation operator is implemented as an integration in time. This process is also called ‘smoothing’. The analysis is repeated as often as necessary for the application, typically with an update frequency of 100–1000 Hz.

14.5 Virtual Source Positioning

Virtual source positioning is a method to control perceived localization using the appropriate reproduction of a monophonic signal. A *virtual source* is an auditory object perceived at a location that does not have any real sources. Virtual source positioning aims to control only the perceived direction of virtual sources, although sometimes the distance and the spatial width of sources may also be controlled.

14.5.1 Amplitude Panning

Amplitude panning means feeding a sound signal $x(t)$ to loudspeakers with different amplitudes, mathematically expressed as

$$x_i(t) = g_i x(t), \quad i = 1, \ldots, N, \tag{14.5}$$

where $x_i(t)$ is the signal fed to loudspeaker i, g_i is the gain of the corresponding channel, N is the number of loudspeakers, and t is time. The listener perceives a virtual source in a direction that depends on the gains. A *panning law* estimates the perceived direction θ from the gains of the loudspeakers. The estimated direction is called the *panning direction* or the *panning angle*. Amplitude panning is conceptually equivalent to the coincident microphone techniques discussed in Section 14.4.3, since the captured signals in these techniques differ, in principle, only in amplitude.

14.5.2 Amplitude Panning in a Stereophonic Set-up

When a virtual source is generated using amplitude panning for the two-channel stereophonic listening set-up, the same sound is applied to the loudspeakers with potentially different amplitudes. The sound arrives from both loudspeakers at both ears, the ipsilateral sound arriving a little earlier than the contralateral, a phenomenon called *cross talk*. A difference of about 0.5 ms in the arrival time at one ear effectively results in a weighted average of the phase of the signals. The *level* differences between the loudspeakers are thus turned, a little surprisingly, into *phase* differences between the ears (Bauer, 1961). This effect is valid only at low frequencies, below about 1 kHz. At high frequencies, above about 2 kHz, the level differences remain as level differences due to the lack of cross talk caused by the shadowing of the head.

The tangent law by Bennett *et al.* (1985) estimates the perceived direction of a virtual source, and it is expressed as

$$\frac{\tan\theta}{\tan\theta_0} = \frac{g_1 - g_2}{g_1 + g_2}, \tag{14.6}$$

which has been found to estimate the direction best in conditions tests in anechoic conditions. Other panning laws also exist and are reviewed by Pulkki (2001b).

The panning laws only determine the ratio between the gains. To prevent an undesirable change in loudness of the virtual source, depending on panning direction, the sum-of-squares of the gains should be normalized:

$$\sum_{n=1}^{N} g_n^p = 1, \tag{14.7}$$

where $p = 2$. This value of p has been found to be the best in real rooms with some reverberation. Depending on listening room acoustics, different normalization rules may be used (Laitinen *et al.*, 2014; Moore, 1990).

The presented analysis is valid only if the loudspeakers are equidistant from the listener and if their angular distance is no larger than about 60°. These criteria define the best listening area in terms of where the virtual sources are localized between the loudspeakers, which, in practice, is only a few tens of centimetres in the left–right direction. When the listener moves outside this area, the virtual source is localized towards the nearest loudspeaker, which emanates a considerable amount of sound. Such erroneous localization occurs due to the precedence effect.

In principle, amplitude panning methods create a comb-filter effect in the ear canal signal spectra, since the same sound arrives from both loudspeakers at each ear with a small time difference. This effect is clearly audible in an anechoic chamber, where it produces a notch in the spectrum between frequencies of 1 kHz and 2 kHz. Fortunately, this effect is not present when listening in a normal room, since the room reverberation mitigates the colouring effect largely (Pulkki, 2001a). The lack of prominent colouring and the relatively robust directional effect provided by amplitude panning are very probably the reasons why it is included in all mixing consoles as a ‘panpot’ control, making it the most widely used technique to position virtual sources.

14.5.3 Amplitude Panning in Horizontal Multi-Channel Loudspeaker Set-ups

In many cases more than two loudspeakers are placed around the listener. Pairwise amplitude panning (Chowning, 1971) is commonly used to position virtual sources with multi-channel set-ups by applying the sound signal only to the loudspeaker pair between which the virtual source lies. Vector base amplitude panning (VBAP) (Pulkki, 2001b) is a commonly used method to formulate pairwise panning. In 2D VBAP, a loudspeaker pair is specified with two vectors. The unit-length vectors $\mathbf{l}_m$ and $\mathbf{l}_n$ point from the listening position to the loudspeakers. The intended direction of the virtual source (panning direction), represented by the unit vector **p**, is expressed as a linear weighted sum of the loudspeaker vectors

$$\mathbf{p} = g_m\mathbf{l}_m + g_n\mathbf{l}_n. \tag{14.8}$$

Here, g_m and g_n are the gain factors of the respective loudspeakers. The gain factors can be computed from

$$\mathbf{g} = \mathbf{p}^{\mathrm{T}}\mathbf{L}_{mn}^{-1}, \tag{14.9}$$

where $\mathbf{g} = [g_m \; g_n]^{\mathrm{T}}$ and $\mathbf{L}_{mn} = [\mathbf{l}_m \; \mathbf{l}_n]$. The calculated factors are used in amplitude panning as gains of the signals applied to the respective loudspeakers after appropriate normalization, say $||\mathbf{g}|| = 1$.

When a virtual source is panned between loudspeakers, the binaural cues are more or less unnatural, since the summing localization mechanism produces somewhat unnatural cues even in the stereophonic case. With loudspeaker pairs not symmetric about the median plane, the produced ITD and ILD cues suggest different directions at different frequencies and are also biased somewhat towards the median plane (Pulkki, 2001b). As a result, the perceived spatial width of the virtual source varies with the panning direction. Thus, the directions of the loudspeakers are perceived if a moving source is created, since the virtual sources are more point-like in their directions. Such an uneven directional width of the sources can be compensated for by blurring the virtual sources slightly in the directions of the loudspeakers. In practice, the sound is always applied to more than one loudspeaker (Pulkki, 2001b; Sadek and Kyriakakis, 2004).

14.5.4 3D Amplitude Panning

A three-dimensional loudspeaker set-up here means a set-up in which the loudspeakers are not all in the same plane. Thus, the set-up has some loudspeakers above and/or below the plane of the horizontal loudspeaker set-up. Triplet-wise panning can be used in such set-ups (Pulkki, 1997), wherein a sound signal is applied to at most three loudspeakers at a time, forming a source triangle from the listener's viewpoint. If more than three loudspeakers are available, the set-up is divided into triangles, one of which is used in the panning of a single virtual source at any one time, as shown in Figure 14.7.

Three-dimensional vector base amplitude panning (3D VBAP) is a method to position virtual sources using such set-ups (Pulkki, 1997). It is formulated similarly to the pairwise panning in the previous section. However, now the gain factors are $\mathbf{g} = [g_m \; g_n \; g_k]^{\mathrm{T}}$, the direction vectors are 3D Cartesian vectors, and the loudspeaker vector base is $\mathbf{L}_{mnk} = [\mathbf{l}_m \; \mathbf{l}_n \; \mathbf{l}_k]$ in

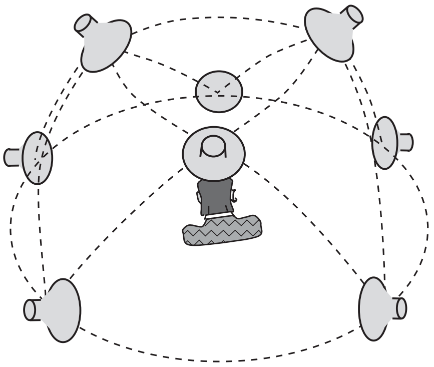

Figure 14.7 A 3D triangulated loudspeaker system for triplet-wise panning.

Equation (14.9). The equation estimates the perceived virtual source direction with relatively good accuracy. The confusion cone azimuth φ_{cc} is estimated to an accuracy of a few degrees in most cases. The perceived confusion cone elevation of a virtual source δ_{cc} is personal for each subject (Pulkki, 2001b), but it is typically perceived inside the loudspeaker triplet.

14.5.5 Virtual Source Positioning using Ambisonics

The Ambisonics microphone technique, discussed in Section 14.4.6, can also be simulated in the computational domain to position virtual sources for 2D or 3D loudspeaker set-ups (Furse, 2009; Malham and Myatt, 1995). The quality issues of Ambisonics as a recording technique, discussed in Section 14.4.6, can be avoided by choosing the order for B-format signals that matches the loudspeaker set-up. The B-format signals produced by a single virtual source can be simulated simply by using such directional patterns as shown in Figure 14.5 ideally, as virtual higher-order microphones do not suffer from self-noise or deformed directional patterns. When multiple concurrent virtual sources exist, their signals can be simply summed to the same B-format audio bus, which corresponds to ideal B-format recording of the corresponding real multi-source scenario. The B-format audio can then be manipulated efficiently. For example, the rotation of the spatial audio scene can be done with simple operations.

14.5.6 Wave Field Synthesis

An alternative goal of sound reproduction is to control the sound field within the loudspeaker set-up. This means that the pressure and velocity waves originating from the virtual sources are to be synthesized as plane waves or as spherical waves. The targeted sound field would then be well defined, and the reproduced sound field can also be simulated, or even measured. Unfortunately, the errors in the synthesis of the sound field between the targeted and reproduced fields do not say much about the perceptual significance of the deviation. Sometimes a small difference in the wave field causes a large perceptual deviation, and a large deviation in the wave field may be perceptually irrelevant in some cases. However, the advantage of this starting point is that the reproduction techniques can be derived mathematically.

Wave field synthesis (WFS) is a technique that requires a large number of carefully equalized loudspeakers (Ahrens, 2012; Berkhout *et al.*, 1993; Vries and Boone, 1999). It aims to reconstruct the entire sound field in a listening room. When a virtual source is reproduced,

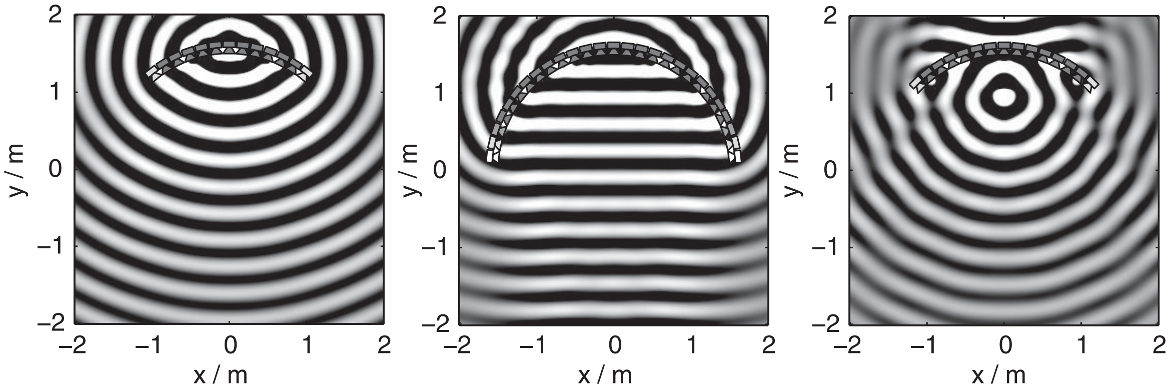

Figure 14.8 The wave field synthesis concept. A circular 56-loudspeaker set-up with radius of 1.5 m is used to reproduce a virtual source emitting a 1 kHz sinusoid 0.5 m behind the set-up (left), a plane wave (centre), and a virtual source 0.5 m inside the set-up (right). The colour of the loudspeaker indicates the level of radiation. Inactive loudspeakers are not shown at all. The illustrations were rendered using the sound field synthesis toolbox (Wierstorf and Spors, 2012).

the sound for each loudspeaker is delayed and amplified so that a desired circular or planar sound wave results as a superposition of sounds from each loudspeaker. The virtual source can be positioned far behind the loudspeakers or, in some cases, even in the space inside the loudspeaker array, as shown in Figure 14.8.

Usually, WFS is used to synthesize virtual sources by specifying the location of the source in the virtual world and using the loudspeaker set-up to generate the wave field. The use of actual recordings for WFS is very rare, probably because setting up the hundreds of carefully equalized microphones required is technically very demanding. This is the reason why WFS is discussed in this book in the section on virtual source positioning and not in the section on microphone techniques. In practice, the microphone techniques discussed in the Section 14.4.5 are used for recording, and the recorded signals are then reproduced with the appropriate system.

Theoretically, the WFS is superior as a technique, because the perceived position of the sound source is correct within a very large listening area. Unfortunately, to create the desired wave field in the total area inside the array requires that the loudspeakers are at a distance of at most half a wavelength from each other. Arrays for wave field synthesis have been built for room acoustics control and enhancement to be used in theatres and multi-purpose auditoria (Vries and Boone, 1999).

Another application of wave field synthesis techniques using a large number of loudspeakers is a method called *sound field control* (Francombe *et al.*, 2013; Nelson and Elliott, 1992). Sound field control is used for various applications. For example, different signals can be made audible in different zones of a large listening area in such a manner that only one of the signals is heard in one zone. The method may also be used to attenuate noise in a large area, for example, to reduce background noise for all passengers in the cabin of an aeroplane.

14.5.7 Time Delay Panning

When the signal to one loudspeaker in a stereophonic set-up is delayed by a constant amount, virtual sources with transient signals are perceived to migrate towards the loudspeaker that radiates the earlier sound signal (Blauert, 1996). The maximal effect is achieved asymptotically when the delay is approximately 1.0 ms or more.

Such processing converts the *phase or time* delays between the loudspeakers at low frequencies to a perception of *level* differences between the ears, and at high frequencies the perception remains as a *time* difference in the signal between the ears. This effect makes the virtual source direction depend on the signal itself (Cooper, 1987; Lipshitz, 1986). The produced binaural cues vary with frequency, and different cues suggest different directions for virtual sources (Pulkki, 2001b). The cue may thus generate a 'spread' perception of the direction of sound, which is desirable in some cases. The effect is also dependent on listening position. For example, if the sound signal is delayed by 1 ms in one loudspeaker, the listener can compensate for the delay by moving a bit towards the delayed loudspeaker. Time delay panning thus resembles spaced microphone techniques, since the spacing between the microphones causes time delays between microphone signals and since the resulting spatial image is similar.

A special case of phase difference in stereophonic reproduction is the use of antiphasic signals in the loudspeakers, where the same signal is applied to both loudspeakers but with the polarity inverted at one loudspeaker, producing a constant 180° phase difference between the signals at all frequencies. This phase difference changes the perceived sound colour, and also spreads the virtual sources. Depending on the listening position, low frequencies may be cancelled out. At higher frequencies this effect is milder, since when the wavelength becomes shorter, the listening area where the loudspeaker signals cancel each other shrinks, and beyond some frequency one of the ears will be outside this listening region where the signals cancel. This effect is also milder in rooms with longer reverberation. The directional perception of the antiphasic virtual source depends on the listening position. In the best listening position, the high frequencies are perceived at the centre and low frequencies in random directions. Outside the best position, the direction is either random or towards the closest loudspeaker. In the language of audio engineers, this effect in the sound is called 'phasy', or 'there is phase error in here'.

14.5.8 Synthesizing the Width of Virtual Sources

A loudspeaker layout with loudspeakers around the listener can also be used to control the width of a virtual source or even to produce an enveloping perception of the sound source. A simple demonstration can be made by playing back pink noise through all loudspeakers independent of each other (Blauert, 1996). The sound source is then perceived to surround the listener completely.

Time–frequency-domain spatial audio processing also provides a means to control the source width effectively. The input signal is divided into frequency channels that are then positioned in different directions around the listener (Pihlajamäki *et al.*, 2014).

14.6 Binaural Techniques

Binaural techniques are loosely defined to be methods that aim to reproduce accurately ear canal signals recorded in a real acoustic scenario with a real subject, or to reproduce ear canal signals which would occur in a virtual world. This is done by recording sound from ear canals or by utilizing measured or modelled head-related transfer functions (HRTFs) and acoustic modelling of listening spaces.

14.6.1 Listening to Binaural Recordings with Headphones

The basic *binaural recording* technique is to reproduce a recorded binaural sound track through headphones. The recording is made by inserting miniature microphones in the ear canals of a real human listener, or by using a manikin with microphones in the ears (Blauert, 1996; Wilska, 1938). Such a recording is reproduced by playing the recorded signals to the ears of the listener. In principle, this is a very simple technique and can provide effective results. A simple implementation is to replace the transducers of in-ear headphones with miniature microphones, use a portable audio recorder to record the sounds of the surroundings, and play back the sound with headphones. Without any further processing, a convincing spatial effect is already achieved, as the left–right directions of the sound sources and the reverberant sound field are reproduced naturally. If the person who did the recording is the listener, the effect can be particularly striking.

Unfortunately, there are also technical challenges with the technique. The sound may appear coloured, the perceived directions move from front to back, and everything may be localized inside the head. To partially avoid these problems, the recording and the reproduction should be carefully equalized, because headphone listening typically produces a different magnitude spectrum to the ear drum than natural listening. Careful equalization of headphone listening is, unfortunately, a complicated business, and it requires very careful measurements of the acoustic transmission of sound from the headphone to the ear drum (Xie, 2013).

A further challenge in binaural reproduction is that the auditory system also utilizes dynamic cues to localize sound sources, as discussed in Section 12.3.4 on page 232. When listening to a binaural recording with headphones, the movements of the listener do not change the binaural reproduction at all. This is one reason why headphone reproduction easily tends to be localized inside the head of the listener (Blauert, 1996).

Another issue is the problem of individuality. Every listener has a unique pinna and head size, and sound in similar conditions seems different in different individuals' ears. When listening to a binaural recording made by another individual, similar problems occur as with non-optimal equalization (Rumsey, 2011; Wenzel *et al.*, 1993).

14.6.2 HRTF Processing for Headphone Listening

A monophonic sound signal can be virtually positioned in any direction in headphone listening if the HRTFs for both ears are available for the desired virtual source direction (Møller *et al.*, 1995; Xie, 2013). A signal x_m, meant to be perceived to be arriving from a certain direction, is convolved with the HRIR pair $\{H_\mathrm{r}, H_\mathrm{l}\}$ measured with the source in the same direction, and the convolved signals

$$y_\mathrm{l} = H_\mathrm{l} \star x_m, \quad \text{and} \tag{14.10}$$

$$y_\mathrm{r} = H_\mathrm{r} \star x_m \tag{14.11}$$

are applied to the headphones, as shown in Figure 14.9a. Since the decay time of the HRIR is always less than a few milliseconds, 256 to 512 taps in the filters are sufficient at a sampling rate of 44.1 kHz. The method ideally reproduces the ear canal signals that would have been produced had the sound source existed in the desired direction. The response of the headphones is assumed to be equalized to be the ideal dirac impulse.

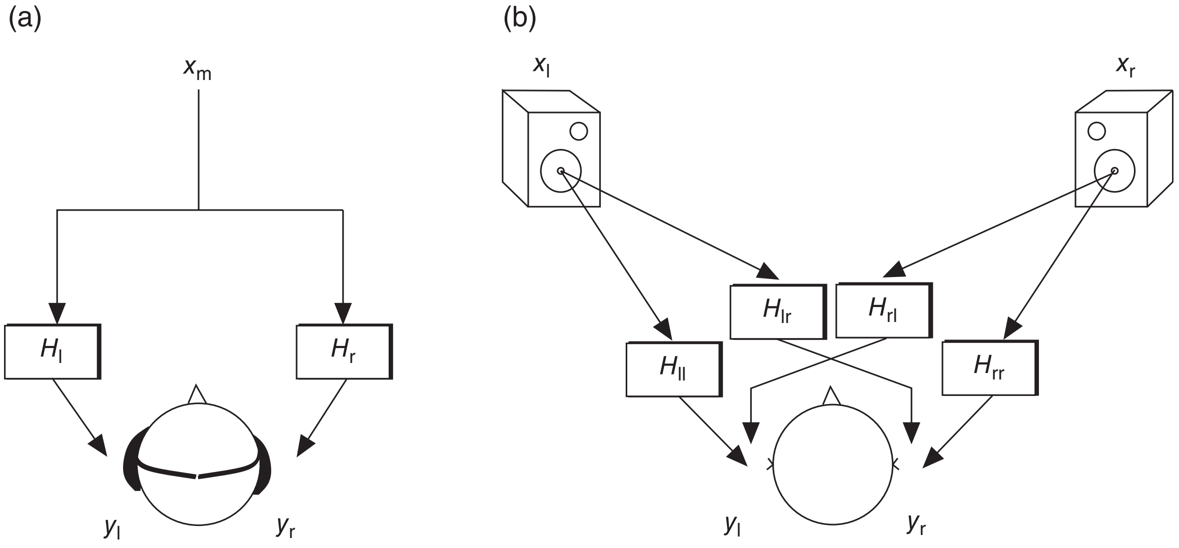

Figure 14.9 (a) Creating a virtual source with HRTF processing. (b) HRTFs in stereophonic listening.

The simple HRTF processing described above very probably produces the perception of an inside-head virtual source, which may also be perceived to be coloured if the headphones are not calibrated carefully. Using a head tracker, a far more realistic perception of external sound sources can be obtained. In *head tracking*, the direction of the listener's head is monitored about 10–100 times a second, and the HRTF filter is changed dynamically to keep the perceived direction of sound constant in the global coordinates (Begault *et al.*, 2001; Breebaart and Schuijers, 2008). In practice, the updating of the HRTF filter has to be done carefully so as not to produce audible artefacts.

The above-discussed technique simulates anechoic listening. It is also possible to simulate the binaural listening of a sound source in a real room. In this approach, the *binaural room impulse responses* (BRIRs) are measured from the ear canals of a subject in a room with a relatively distant source (Blauert, 1996; Møller, 1992). The response thus contains the contributions of the direct sound, reflections, and reverberation. When the response is convolved with a signal and reproduced over headphones, listeners would optimally perceive the sound as if they were in the room where the BRIR was measured. Additionally, the presence of the room response in the reproduction may cause externalization of the perceived auditory scene.

14.6.3 Virtual Listening of Loudspeakers with Headphones

An interesting application for HRTF technologies with headphones is listening to of existing multi-channel audio content. Here, each loudspeaker of a multi-channel loudspeaker layout is simulated using an HRTF set. For example, to listen to stereophonic signals $\{x_l, x_r\}$ over virtual loudspeakers in the directions of $\{-30°, 30°\}$, the signals are convolved with the HRTFs for each loudspeaker measured from the corresponding directions, as shown in Figure 14.9b (Blauert, 1996; Kirkeby, 2002). The convolved signals

$$y_l = H_{ll} \star x_l + H_{rl} \star x_r, \quad \text{and} \tag{14.12}$$

$$y_r = H_{rr} \star x_r + H_{lr} \star x_l \tag{14.13}$$

are applied to the headphones. The headphones are not shown in the figure.

The use of HRTFs measured in anechoic conditions is problematic in many cases. In most cases the listener is situated in a normal room or outdoors, and reproducing anechoic binaural sound to the listener may cause the virtual sources to be localized inside the head. The use of BRIRs, which optimally have a similar response as in the room where the listener is located is beneficial here. Correcting the effect of the HRTFs measured in anechoic conditions can also be done using measured room impulse responses, or by simulating the effect or room with a reverberator (Mackensen *et al.*, 1999).

14.6.4 Headphone Listening to Two-Channel Stereophonic Content

Headphone listening to stereophonic content without HRTF processing is significantly different from loudspeaker listening to the same stereophonic content. The cross talk present in loudspeaker listening is missing in headphone reproduction meaning that the sound from the left headphone only enters the left ear canal and likewise on the right side. Typically, audio engineers create the stereophonic audio content in studios with two-channel loudspeaker listening. A relevant question, then, is how does the spatial perception of the content change when listened to with headphones?

With amplitude-panned virtual sources the level difference between headphone channels is converted directly into an ILD, and the ITD remains zero. This is very different from loudspeaker listening, where the direction of the amplitude-panned sources is implied by ITD cues with the ILD at zero for low frequencies. Although this appears to be a potential source of large differences in spatial perception of the resulting virtual sources, the resulting spatial image is similar. The virtual sources are perceived in about the same order in the left–right direction as in loudspeaker listening. However, in headphone listening, the sources are perceived inside the listener's head due to erroneous monaural spectra and a lack of dynamic cues, as explained previously.

If the stereophonic content includes virtual sources that have been reproduced with time delays between the loudspeaker channels, as explained in Section 14.5.7, the result may be a vastly different spatial perception in headphone listening. For example, a 3-ms delay in the left loudspeaker may produce spread perception of the sound in loudspeaker listening, but in headphone listening the sound most probably is perceived to originate only from the right headphone.

14.6.5 Binaural Techniques with Cross-Talk-Cancelled Loudspeakers

Binaural recordings are meant to be played back in such a manner that the sound originating at the left ear should be applied to the left ear only, and correspondingly with the right ear, as depicted in Figure 14.9a. If such a recording is played back with a stereophonic set-up of loudspeakers, the sound from the left loudspeaker also enters the right ear, and vice versa, as shown in Figure 14.9b. This mixing of signals is called *cross talk*, which should be avoided in binaural playback.

In order to be able to listen to binaural recordings over two loudspeakers, methods to cancel cross talk have been proposed (Cooper and Bauck, 1989; Kirkeby *et al.*, 1998; Mouchtaris *et al.*, 2000). A system can be built to deliver binaurally recorded signals to the listener's ears using two closely spaced loudspeakers with crosstalk cancellation. The binaural signals are represented as a 2×1 vector $\mathbf{x}(n)$ and the produced ear canal signals also as a 2×1 vector $\mathbf{d}(n)$. The system can be described in the Z domain as

$$\mathbf{d}(z) = \mathbf{H}(z)\mathbf{G}(z)\mathbf{x}(z), \tag{14.14}$$

where $\mathbf{H}(z) = \begin{bmatrix} H_{\mathrm{ll}}(z) & H_{\mathrm{lr}}(z) \\ H_{\mathrm{rl}}(z) & H_{\mathrm{rr}}(z) \end{bmatrix}$ contains the electro-acoustic responses of the loudspeakers measured in the ear canals, as shown in the figure, and $\mathbf{G}(z) = \begin{bmatrix} G_{\mathrm{ll}}(z) & G_{\mathrm{lr}}(z) \\ G_{\mathrm{rl}}(z) & G_{\mathrm{rr}}(z) \end{bmatrix}$ contains the responses for performing inverse filtering to minimize the cross talk.

Ideally, $\mathbf{x}(z) = \mathbf{d}(z)$, which is obtained if $\mathbf{G}(z) = \mathbf{H}(z)^{-1}$. Unfortunately, direct inversion of the matrix is not feasible due to the non-idealities of the loudspeakers and the listening conditions. A regularized method to find an optimal $\mathbf{G}_{\mathrm{opt}}(z)$ has been proposed by Kirkeby *et al.* (1998):

$$\mathbf{G}_{\mathrm{opt}}(z) = \left[\mathbf{H}^{\mathrm{T}}(z^{-1})\mathbf{H}(z) + \beta\mathbf{I}\right]^{-1} \mathbf{H}^{\mathrm{T}}(z^{-1})z^{-m}, \tag{14.15}$$

where β is a positive scalar regularization factor and z^{-m} models the time delay due to the sound reproduction system. If β is selected to be very small, sharp peaks will result in the time-domain inverse filters, which may exceed the dynamic range of the loudspeakers. If β is larger, the impulse response of the inverse filter will have a longer duration in time, which is less demanding on the loudspeakers, but the price paid is that the inversion is less accurate (Kirkeby *et al.*, 1998).

In practice, this method performs best with loudspeakers near each other, since a larger loudspeaker base angle leads to colouration at lower frequencies. The listening area where the effect is audible is very small, because if the listener departs from the mid-line between the loudspeakers, a region about 1–2 cm wide, the effect is lost.

A nice feature of this technique is that the sound is typically externalized. This may be because head movements of the listener produce somewhat relevant cues, and because the sound is reproduced using far-field loudspeakers generating correct monaural directional spectral cues. However, although the sound is externalized, it is hard to reproduce the virtual sources in all directions with this technique. With a stereo dipole in the front, the reproduced sound scene is typically perceived only in the frontal hemisphere.

The technique is also sensitive to the reflections and reverberation of the listening room. It performs at its best only in spaces without prominent reflections. To obtain the best results, the HRTFs of the listener should be known, but very plausible results can be obtained with generic responses.

14.7 Digital Audio Effects

Digital audio effects are systems that modify audio signals fed at their inputs according to set control parameters and make the modified signal available at their outputs (Zölzer, 2011). The purpose of the processing is to modify perceptual characteristics of sound to meet artistic needs in audio engineering. The settings of the parameters of the effects are made by audio engineers, musicians, and even by the listeners of music or other audio. Some of the techniques already described in this book, such as equalization and virtual source positioning, can be classified as audio effects. Other examples of audio effects are:

- *Dynamic range control*, *dynamic processing*, or *automatic gain control*: The effect comprises a gain that is automatically controlled by the level of the input signal (Zölzer, 2008)

(see Figure 19.10 on page 412). A *dynamic range controller*, a.k.a. a *compressor*, attenuates sound the more the level increases beyond a certain threshold. A *limiter* is a similar effect, but limits the level of signal to a preset value. An *expander*, on the other hand, leaves large amplitudes untouched, but increasingly amplifies the signal the smaller the amplitude is. These techniques will be discussed later in this section, and also in the context of hearing aids in Section 19.5.2.
- *Pitch shifting*: The pitch of harmonic complexes can be shifted with different methods. A basic method is to re-sample the signal with time-scaling, which also scales the magnitude spectrum up or down. There are many methods to change only the pitch, leaving the spectral characteristics unchanged (Bristow-Johnson, 1995; Moulines and Laroche, 1995). This effect enables, for example, the correction of out-of-tune singing or the creation of different pitches from a sampled note of a musical instrument.
- *Chorus*, *flanger*, and *phaser*: These are effects where at least one copy of the original signal is modulated or changed in phase or pitch and the modified signal(s) is added to the original signal. This effectively cancels and enhances partials of the original signal and/or changes somewhat the spectral structure of the original signal (Orfanidis, 1995). For example, when the spectral structure of a single voice is smeared slightly in frequency, the processed voice may resemble the sound of a choir, which is the principle of the *chorus* effect.
- *Room effects*: These effects simulate the effect of a room on audio signals and are discussed in the next section.

The interested reader is referred to Zölzer (2011) for a more complete list of audio effects.

The dynamic range control methods have evoked a lively discussion among audio engineers – the *loudness war* (Vickers, 2011). A basic purpose of compression is to make loudness evoked by processed sound events roughly constant, and often also to maximize the loudness of the piece of audio content delivered. This processing is advantageous when audio content should be audible over loud background noise existing in the listening conditions. This is typically the case in, say, car audio. A downside of this processing is that the non-linear processing changes the spectrum of the input signal, typically generating harmonic distortion in the output. A problem arises when several compressors are in the signal path: the resulting attributes of perceived audio can differ immensely from those in the mastering studio and may lead to large changes in perceptual attributes of the reproduced sound. This situation can exist in broadcasting, because most broadcasters compress all broadcast audio regardless of whether the content has already been compressed or not; this is possibly followed by further compression by audio devices like those in a car, which also compress their output.

14.8 Reverberators

Generating the effect of reverberation in the produced sound is often desirable in audio content production. A digital audio effect device called a *reverberator* is able to *reverberate* the signal, causing a human listener to perceive the sound to be *reverberant*.

An intuitive method of creating such a room effect is to measure the impulse response of a real room and to compute the convolution of the signal and the impulse response. An alternative approach to obtaining the impulse response of a room is to simulate its acoustics using computational models of room acoustics (see Section 2.4.5). Unfortunately, convolution-based reverberation is not feasible in most cases due to the computational complexity of the filtering.

Instead of convolution, computationally less demanding DSP structures are often used to create the perception of reverberation. There are several methods of obtaining a room effect to make the processed sound reverberant, although the processing itself does not mimic the physical propagation of sound in rooms. These approaches are briefly summarized below.

14.8.1 Using Room Impulse Responses in Reverberators

If the impulse response of a target room is readily available, the most faithful reverberation method is to convolve the input signal with the impulse response. The result corresponds to the imaginary case where the input signal to be reverberated is applied to the loudspeaker in the measurement room and the listener listens to the signal recorded using the microphone used in the measurement. Note that the directional responses of the microphone and the sound source used in the measurement affect the result. For example, the reverberation effect will be considerably different if the directivity of the microphone is changed. An omnidirectional microphone captures the reverberation from all directions equally, while a directional microphone facing towards the source significantly suppresses sound arriving from the reverberant field. If these responses are used in convolving reverberators, the omnidirectional response will be perceived as much more reverberant than the directional response.

Direct convolution can be implemented by storing each sample of the impulse response as a coefficient of an FIR filter whose input is the signal recorded in a free field. Direct convolution easily becomes impractical if the length of the target response exceeds small fractions of a second, since it would translate into several hundreds of taps in the filter structure. A solution is to perform the convolution block by block in the frequency domain. Given the Fourier transforms of the impulse response and of a block of the input signal, the two can be multiplied point by point and the result transformed back to the time domain. As this kind of processing is performed on successive blocks of the input signal, the output signal is obtained by overlapping and adding the partial results (Oppenheim and Schafer, 1975). Thanks to the FFT computation of the discrete Fourier transform, this technique is significantly faster. A drawback is that, in order to be operated in real time, a block of N samples must be read and then processed while a second block is being read. Therefore, the input–output latency in samples is twice the size of a block, and this is not tolerable in practical real-time environments.

When a monophonic signal is convolved with a monophonic response, the room response is perceived to be in the direction of the loudspeaker. In reproduction with multiple loudspeakers, localizing the reverberation to the directions reproducible with the loudspeaker set-up is often desired. The microphone techniques discussed in Section 14.4 can be used for impulse response recording, and their pros and cons also hold in impulse-response-based reverberation. Some dedicated non-linear techniques for impulse response reproduction for multi-channel playback have also been proposed (Farina and Ugolotti, 1999; Merimaa and Pulkki, 2005; Tervo *et al.*, 2013), which overcome some flaws in the microphone techniques. In the non-linear techniques, measured impulse responses are processed into impulse responses for each loudspeaker in multi-channel layouts using instantaneous spatial analysis from the measured response. Although good results have already been obtained, the technical requirements for authentic spatial reproduction of impulse responses with such techniques are an on-going research topic.

The room effect, that is, the room impulse response, can be created by modelling how sound propagates and reflects from surfaces if the geometry of the room is known. This process is called room acoustics modelling, and several techniques are discussed in Section 2.4.5. When the impulse response is computed with such a model and used to reverberate sound, the process is often referred to as *auralization*, meaning making a room effect audible. Unfortunately, the impulse responses computed from models often do not create natural-sounding reverberation. The frequency range of the modelling may not be sufficient, and the late reverberant tail may sound unrealistically bright. In some cases, only the early reflections are modelled accurately, and the late response is generated using DSP structures (Savioja *et al.*, 1999).

14.8.2 DSP Structures for Reverberators

In the second half of the twentieth century, several engineers and acousticians tried to invent electronic devices capable of simulating the long-term effects of sound propagation in enclosures. The most important pioneering work in the field of *artificial reverberation* was that of Manfred Schroeder at the Bell Laboratories in the early 1960s (Schroeder, 1970). Schroeder introduced the recursive *comb filters* and the delay-based *all-pass filters* as computational structures suitable for the inexpensive simulation of complex patterns of echoes. In particular, the all-pass filter based on the recursive delay line has the form

$$y(n) = -g \cdot x(n) + x(n-m) + g \cdot y(n-m), \tag{14.16}$$

where m is the length of the delay in samples. The filter structure is depicted in Figure 14.10, where $A(z)$ is usually replaced by a delay line. The filter results in a dense impulse response and a flat frequency response. The structure rapidly became a standard component used in almost all the artificial reverberators designed until today (Moorer, 1979). It is usually assumed that all-pass filters do not introduce colouration in the input sound. However, this assumption is valid from a perceptual point of view only if the delay line is much shorter than the integration time of the ear, which is about 50–100 ms. If this is not the case, the time-domain effects become much more relevant, and the timbre of the incoming signal is significantly affected.

The shortcomings of such simple methods have been worked on in later generations of reverberators. There exists a vast number of different DSP structures proposed to create the effect of room reverberation (see Välimäki *et al.* 2012). The best reverberators are able to deliver natural-sounding room effects with significantly lower computational complexity than FIR-based reverberators. An advantage of such reverberators is the simpler control of the perceptual

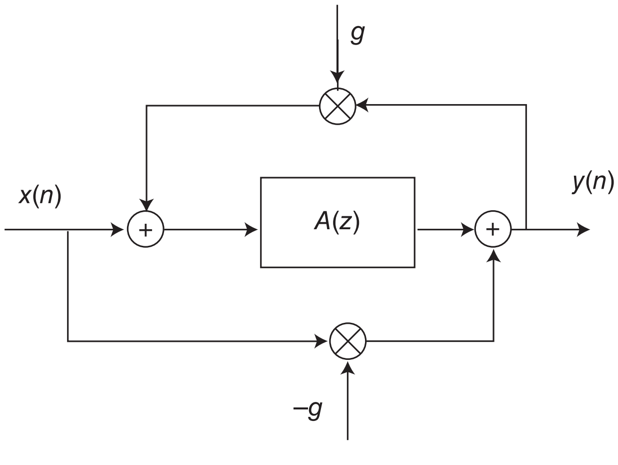

Figure 14.10 The all-pass filter structure.

attributes of the reverberation effect, which can be affected by changing the few parameters of the reverberator. In contrast, to change the parameters of an FIR-based reverberator requires the computation of the modified response and also changing the FIR filter coefficients on the fly, which is less flexible by default.

Summary

Many applications exist where sound is captured, processed, and reproduced. Perhaps the best-known application is audio content production, which serves the cinema, music, and gaming industries. The capturing of sound is performed with microphone techniques designed for different listening set-ups, which may be associated with visual or tactile displays. The microphone techniques have a major effect on the spatial characteristics of the reproduced sound, which can also be emulated using virtual source positioning techniques. Some shortcomings of traditional microphone techniques can be circumvented using non-linear time–frequency-domain reproduction techniques that have been developed by exploiting the knowledge of human spatial hearing resolution. The recorded sounds may also be processed using audio effects to modify the perceptual properties.

Further Reading and Available Toolboxes

The reproduction of sound is a wide topic with many directions in which the interested reader may continue his or her studies. Audio engineering and microphone techniques are covered by Katz and Katz (2007), Owsinski and O'Brien (2006), and Toole (2012). The time frequency-domain techniques for spatial sound reproduction are elaborated in Ahonen (2013), Laitinen (2014), and Vilkamo (2014). A concise summary of recent activities in binaural technologies is made Xie (2013). Digital audio effects can be studied further in Pirkle (2012), Reiss and McPherson (2014), and Zölzer (2011).

A number of software packages exist to perform spatial sound synthesis:

- *SoundScape Renderer* (TU Berlin/Universität Rostock). WFS, VBAP, Ambisonics, and binaural synthesis:
  `http://spatialaudio.net/ssr/`.
- *Panorama* (WaveArts/William Gardner). Binaural rendering with room modelling, and cross talk cancellation:
  `http://wavearts.com/products/plugins/panorama/`
- *Ambdec* (Fons Andriensen). Traditional, high-order, and optimized Ambisonics:
  `http://kokkinizita.linuxaudio.org/linuxaudio/index.html`
- *Blue Ripple Sound* (Richard Furse). Traditional and higher-order Ambisonics, HRTF rendering, and cross talk cancellation:
  `http://www.blueripplesound.com/product-listings/pro-audio`
- *Harpex* (Svein Berge). Time–frequency-domain parametric reproduction of B-format signals:
  `http://www.harpex.net/`
- *VBAP* (Aalto University). VBAP for 2D and 3D loudspeaker layouts:
  `http://www.acoustics.hut.fi/software/vbap/`
- *SPAT* (IRCAM). Panning, Ambisonics, room acoustics modelling, and sound diffusion:
  `http://www.fluxhome.com/products/plug_ins/ircam_spat`

- *Ambitools* (NTNU/Peter Svensson). Higher-order Ambisonics: `http://www.iet.ntnu.no/s̃vensson/software/index.html#AMBI`
- *sWonder* (Marije Baalman). WFS and binaural synthesis: `http://sourceforge.net/projects/swonder/`

References

Adelstein, B.D., Begault, D.R., Anderson, M.R., and Wenzel, E.M. (2003) Sensitivity to haptic–audio asynchrony *Proc. 5th Int. Conf. on Multimodal Interfaces*, pp. 73–76 ACM.

Ahonen, J. (2013) *Microphone front-ends for spatial sound analysis and synthesis with Directional Audio Coding*. PhD thesis, Aalto University.

Ahrens, J. (2012) *Analytic Methods of Sound Field Synthesis*. Springer.

Alexandridis, A., Griffin, A., and Mouchtaris, A. (2013) Capturing and reproducing spatial audio based on a circular microphone array. *J. Elec. Comp. Eng.* **2013**, 7.

Altinsoy, M.E. and Merchel, S. (2010) Cross-modal frequency matching: Sound and whole-body vibration. In Nordahl, R., Serafin, S., Fontana, F., and Brewster, S. (eds) *Haptic and Audio Interaction Design*. Springer, pp. 37–45.

Bauer, B.B. (1961) Phasor analysis of some stereophonic phenomena. *J. Acoust. Soc. Am.*, **33**, 1536–1539.

Bech, S. and Zacharov, N. (2006) *Perceptual Audio Evaluation – Theory, Method and Application*. John Wiley & Sons.

Begault, D.R., Wenzel, E.M., and Anderson, M.R. (2001) Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source. *J. Audio Eng. Soc.*, **49**(10), 904–916.

Bennett, J.C., Barker, K., and Edeko, F.O. (1985) A new approach to the assessment of stereophonic sound system performance. *J. Audio Eng. Soc.*, **33**(5), 314–321.

Berge, S. and Barrett, N. (2010) A new method for B-format to binaural transcoding. *40th Int. Audio Eng. Soc. Conf.: Spatial Audio*. AES.

Berkhout, A., Vries, D., and Vogel, P. (1993) Acoustics control by wave field synthesis. *J. Acoust. Soc. Am.*, **93**(5), 2764–2778.

Björk, S. and Holopainen, J. (2005) *Patterns in Game Design*. Cengage Learning.

Blauert, J. (1996) *Spatial Hearing – Psychophysics of Human Sound Localization*. MIT Press.

Blumlein, A.D. (1931) U.K. Patent 394,325. Reprinted in 1986 *Stereophonic Techniques*, AES.

Borenius, J. (1985) Perceptibility of direction and time delay errors in subwoofer reproduction. *79th Audio Eng. Soc. Convention*. AES.

Breebaart, J. and Schuijers, E. (2008) Phantom materialization: A novel method to enhance stereo audio reproduction on headphones. *IEEE Trans. Audio, Speech, and Language Proc.*, **16**(8), 1503–1511.

Bristow-Johnson, R. (1995) A detailed analysis of a time-domain formant-corrected pitch-shifting algorithm. *J. Audio Eng. Soc.*, **43**(5), 340–352.

BS.775-2, I. (2006) Multichannel stereophonic sound system with and without accompanying picture. Recommendation, International Telecommunication Union, Geneva, Switzerland.

Chowning, J. (1971) The simulation of moving sound sources. *J. Audio Eng. Soc.*, **19**(1), 2–6.

Cobos, M., Lopez, J., and Spors, S. (2010) A sparsity-based approach to 3D binaural sound synthesis using time–frequency array processing. *EURASIP J. Adv. Sig. Proc.* **2010**, 1–13.

Cooper, D.H. (1987) Problems with shadowless stereo theory: Asymptotic spectral status. *J. Audio Eng. Soc.*, **35**(9), 629–642.

Cooper, D.H. and Bauck, J.L. (1989) Prospects for transaural recording. *J. Audio Eng. Soc.*, **37**(1/2), 3–39.

Coutrot, A., Guyader, N., Ionescu, G., and Caplier, A. (2012) Influence of soundtrack on eye movements during video exploration. *J. Eye Movement Res.*, **5**(4), 1–10.

Davis, M.F. (2003) History of spatial coding. *J. Audio Eng. Soc.*, **51**(6), 554–569.

Farina, A. and Ugolotti, E. (1999) Subjective comparison between stereo dipole and 3D ambisonic surround systems for automotive applications. *16th Int. Audio Eng. Soc. Conf: Spatial Sound Reproduction*. AES.

Francombe, J., Coleman, P., Olik, M., Baykaner, K., Jackson, P.J., Mason, R., Dewhirst, M., Bech, S., and Pederson, J.A. (2013) Perceptually optimized loudspeaker selection for the creation of personal sound zones. *52nd Int. Audio Eng. Soc. Conf.: Sound Field Control-Engineering and Perception*. AES.

Furse, R.W. (2009) Building an open AL implementation using ambisonics. *35th Int. Audio Eng. Soc. Conf.: Audio for Games* AES.

Gerzon, M.J. (1973) Periphony: With height sound reproduction. *J. Audio Eng. Soc.*, **21**(1), 2–10.

Hamasaki, K. and Hiyama, K. (2003) Reproducing spatial impression with multichannel audio. *24th Int. Conf. Audio Eng. Soc.: Multichannel Audio, The New Reality*. AES.

Hamasaki, K., Hiyama, K., and Okumura, R. (2005) The 22.2 multichannel sound system and its application. *Audio Eng. Soc. Convention 118* AES.

Hollier, M.P., Rimell, A.N., Hands, D.S., and Voelcker, R.M. (1999) Multi-modal perception. *BT Technol. J.*, **17**(1), 35–46.

ISO/IEC 23008-1 (2014) High efficiency coding and media delivery in heterogeneous environments. Standard.

ITU-R BS.1116-1 (1997) Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems. Recommendation, International Telecommunication Union, Geneva, Switzerland.

ITU-T (1990) Tolerances for transmission time differences between the vision and sound components of a television signal. Recommendation J.100, International Telecommunication Union, Geneva, Switzerland.

Joly, A., Montard, N., and Buttin, M. (2001) Audio-visual quality and interactions between television audio and video *Sixth International Symposium on Signal Processing and its Applications 2001*, volume 2, pp. 438–441 IEEE.

Katz, B. and Katz, R.A. (2007) *Mastering Audio: The art and the science*. Taylor & Francis US.

Kirkeby, O. (2002) A balanced stereo widening network for headphones. *22nd Int. Audio Eng. Soc. Conf.: Virtual, Synthetic, and Entertainment Audio* AES.

Kirkeby, O., Nelson, P., and Hamada, H. (1998) Local sound field reproduction using two closely spaced loudspeakers. *J. Acoust. Soc. Am.*, **104**, 1973–1981.

Kohlrausch, A. and van de Par, S. (2005) Audio-visual interaction in the context of multi-media applications. In Blauert, J. (ed.) *Communication Acoustics*. Springer, pp. 109–138.

Kyriakakis, C. (1998) Fundamental and technological limitations of immersive audio systems. *Proc. of the IEEE*, **86**(5), 941–951.

Laitinen, M-V., (2014) Techniques for versatile spatial-audio reproduction in time-frequency domain. PhD thesis, Aalto University.

Laitinen, M-V., Vilkamo, J., Jussila, K., Politis, A., and Pulkki, V. (2014) Gain normalization in amplitude panning as a function of frequency and room reverberance. *55th Int. Audio Eng. Soc. Conf.: Spatial Audio*. AES.

Lemieux, P.A.S., Dressler, W., and Jot, J.M. (2013) Object-based audio system using vector base amplitude panning. US Patent App. 13/906,214.

Levitin, D.J., MacLean, K., Mathews, M., Chu, L., and Jensen, E. (2000) The perception of cross-modal simultaneity (or 'the Greenwich observatory problem' revisited). *AIP Conference Proceedings*, volume 517, pp. 323–329.

Lipshitz, S.P. (1986) Stereophonic microphone techniques... are the purists wrong? *J. Audio Eng. Soc.*, **34**(9), 716–744.

Mackensen, P., Felderhoff, U., Theile, G., Horbach, U., and Pellegrini, R.S. (1999) Binaural room scanning–A new tool for acoustic and psychoacoustic research. *J. Acoust. Soc. Am.*, **105**(2), 1343–1344.

Malham, D.G. and Myatt, A. (1995) 3-D sound spatialization using ambisonic techniques. *Comp. Music J.*, **19**(4), 58–70.

Menzel, D., Fastl, H., Graf, R., and Hellbrück, J. (2008) Influence of vehicle color on loudness judgments. *J. Acoust. Soc. Am.*, **123**, 2477–2479.

Merchel, S. and Altinsoy, M.E. (2013) Vibration in music perception. *Audio Eng. Soc. Convention 134* AES.

Merchel, S., Altinsoy, E., and Stamm, M. (2010) Tactile music instrument recognition for audio mixers. *Audio Eng. Soc. Convention 128* AES.

Merchel, S., Leppin, A., and Altinsoy, E. (2009) Hearing with your body: the influence of whole-body vibrations on loudness perception. *16th Int. Conf. Sound and Vibration*.

Merimaa, J. and Pulkki, V. (2004) Spatial impulse response rendering. *7th Intl. Conf. on Digital Audio Effects (DAFXÕ04)*.

Merimaa, J. and Pulkki, V. (2005) Spatial impulse response rendering I: Analysis and synthesis. *J Audio Eng. Soc.*, **53**(12), 1115–1127.

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

Møller, H., Sørensen, M.F., Hammershøi, D., and Jensen, C.B. (1995) Head-related transfer functions of human subjects. *J. Audio Eng. Soc.*, **43**(5), 300–321.

Moore, F.R. (1990) *Elements of Computer Music*. Prentice Hall.

Moorer, J.A. (1979) About this reverberation business. *Computer Music J.*, **3**(2), 13–28.

Moreau, S., Daniel, J., and Bertet, S. (2006) 3D sound field recording with higher order Ambisonics – objective measurements and validation of spherical microphone. *Audio Eng. Soc. Convention 120*. AES.

Mouchtaris, A., Reveliotis, P., and Kyriakakis, C. (2000) Inverse filter design for immersive audio rendering over loudspeakers. *IEEE Trans. Multimedia*, **2**(2), 77–87.

Moulines, E. and Laroche, J. (1995) Non-parametric techniques for pitch-scale and time-scale modification of speech. *Speech Commun.* **16**(2), 175–205.

Nelson, P. and Elliott, S. (1992) *Active Control of Sound*. Academic Press.

Oppenheim, A.V. and Schafer, R.W. (1975) *Digital Signal Processing*. Prentice-Hall.

Orfanidis, S.J. (1995) *Introduction to Signal Processing*. Prentice-Hall,

Owsinski, B. and O'Brien, M. (2006) *The Mixing Engineer's Handbook*. Thomson Course Technology.

Pätynen, J. and Lokki, T. (2010) Directivities of symphony orchestra instruments. *Acta Acustica United with Acustica*, **96**(1), 138–167.

Pihlajamäki, T., Santala, O., and Pulkki, V. (2014) Synthesis of spatially extended virtual sources with time–frequency decomposition of mono signals. *J. Audio Eng. Soc.* **62**(7/8), 467–484.

Pirkle, W. (2012) *Designing Audio Effect Plug-Ins in C++: With Digital Audio Signal Processing Theory*. Taylor & Francis.

Politis, A., Vilkamo, J., and Pulkki, V. n.d. Sector-based perceptual sound field reproduction in the spherical harmonic domain. Unpublished manuscript.

Pulkki, V. (1997) Virtual source positioning using vector base amplitude panning. *J. Audio Eng. Soc.*, **45**(6), 456–466.

Pulkki, V. (2001a) Coloration of amplitude-panned virtual sources. *Audio Eng. Soc. Convention 110*. AES.

Pulkki, V. (2001b) *Spatial Sound Generation and Perception by Amplitude Panning Techniques*. PhD thesis, Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing.

Pulkki, V. (2002) Microphone techniques and directional quality of sound reproduction. *Audio Eng. Soc. Convention 112* AES.

Pulkki, V. (2007) Spatial sound reproduction with directional audio coding. *J. Audio Eng. Soc.*, **55**(6), 503–516.

Rafaely, B., Weiss, B., and Bachmat, E. (2007) Spatial aliasing in spherical microphone arrays. *IEEE Trans. Signal Proc.*, **55**(3), 1003–1010.

Reiss, J.D. and McPherson, A.P. (2014) *Audio Effects: Theory, Implementation and Application*. Taylor & Francis/CRC Press.

Rimell, A. and Owen, A. (2000) The effect of focused attention on audio-visual quality perception with applications in multi-modal codec design. *IEEE Int. Conf. Acoustics, Speech, and Signal Proc.*, volume 6, pp. 2377–2380 IEEE.

Robinson, C.Q., Mehta, S., and Tsingos, N. (2012) Scalable format and tools to extend the possibilities of cinema audio. *SMPTE Motion Imag. J.* **121**(8), 63–69.

Rumsey, F. (2001) *Spatial Audio*. Taylor & Francis.

Rumsey, F. (2006) *Sound and Recording: An introduction*. Taylor & Francis US.

Rumsey, F. (2011) Whose head is it anyway? Optimizing binaural audio. *J. Audio Eng. Soc.*, **59**(9), 672–675.

Sadek, R. and Kyriakakis, C. (2004) A novel multichannel panning method for standard and arbitrary loudspeaker configurations. *Audio Eng. Soc. 117th Convention*, AES.

Savioja, L., Huopaniemi, J., Lokki, T., and Väänänen, R. (1999) Creating interactive virtual acoustic environments. *J. Audio Eng. Soc.*, **47**(9), 675–705.

Schroeder, M. (1970) Digital simulation of sound transmission in reverberant spaces. *J. Acoust. Soc. Am.*, **47**(2), 424–431.

Silzle, A., George, S., Habets, E., and Bachmann, T. (2011) Investigation on the quality of 3D sound reproduction. *Proceedings of ICSA*, p. 334.

Simon, G., Olive, S., and Welti, T. (2009) The effect of whole-body vibration on preferred bass equalization in automotive audio systems. *Audio Eng. Soc. Convention 127* AES.

Solvang, A. (2008) Spectral impairment of two-dimensional higher order Ambisonics. *J. Audio Eng. Soc.*, **56**(4), 267–279.

Steinke, G. (1996) Surround sound – the new phase. An overview *Audio Eng. Soc. 100th Convention*, AES.

Streicher, R. and Dooley, W. (1985) Basic stereo microphone perspectives – a review. *J. Audio Eng. Soc.*, **33**(7/8), 548–556.

Tervo, S., Pätynen, J., Kuusinen, A., and Lokki, T. (2013) Spatial decomposition method for room impulse responses. *J. Audio Eng. Soc.*, **61**(1/2), 17–28.

Toole, F. (2012) *Sound Reproduction: the Acoustics and Psychoacoustics of Loudspeakers and Rooms*. Focal Press.

Torick, E. (1998) Highlights in the history of multichannel sound. *J. Audio Eng. Soc.*, **46**(1/2), 27–31.

Tournery, C., Faller, C., Kuech, F., and Herre, J. (2010) Converting stereo microphone signals directly to MPEG-surround. *Audio Eng. Soc. Convention 128* AES.

Välimäki, V., Parker, J.D., Savioja, L., Smith, J.O., and Abel, J.S. (2012) Fifty years of artificial reverberation. *IEEE Trans. Audio, Speech, and Language Proc.*, **20**(5), 1421–1448.

Vickers, E. (2011) The loudness war: Do louder, hypercompressed recordings sell better? *J. Audio Eng. Soc.*, **59**(5), 346–351.

Vilkamo, J. (2014) *Perceptually Motivated Time–Frequency Processing of Spatial Audio*. PhD thesis, Aalto University.

Vilkamo, J. and Pulkki, V. (2013) Minimization of decorrelator artifacts in directional audio coding by covariance domain rendering. *J. Audio Eng. Soc.*, **61**(9), 637–646.

Vilkamo, J., Lokki, T., and Pulkki, V. (2009) Directional audio coding: Virtual microphone-based synthesis and subjective evaluation. *J. Audio Eng. Soc.*, **57**(9), 709–724.

Vries, D. and Boone, M. (1999) Wave field synthesis and analysis using array technology. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 15–18, Mohonk Mountain House, New Paltz.

Welti, T. (2004) Subjective comparison of single channel versus two channel subwoofer reproduction. *117th Audio Eng. Soc. Convention*. AES.

Wenzel, E.M., Arruda, M., Kistler, D.J., and Wightman, F.L. (1993) Localization using nonindividualized head-related transfer functions. *J. Acoust. Soc. Am.*, **94**(1), 111–123.

Wierstorf, H. and Spors, S. (2012) Sound field synthesis toolbox. *Audio Engineering Society Convention 132* AES.

Wilska, A. (1938) *Untersuchungen über das richtungshoeren (Studies on Directional Hearing)*. PhD thesis, Helsinki University. English translation available: `http://www.acoustics.hut.fi/publications/Wilskathesis/`.

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

Zölzer, U. (2008) *Digital Audio Signal Processing*. John Wiley & Sons.

Zölzer, U. (2011) *DAFX: Digital Audio Effects*. John Wiley & Sons.