Chapter 2
Literature Review on Spatial Audio

Abstract Inspired by the human auditory system, the sound scene is considered as the mixture of a foreground sound (primary component, directional) and a background sound (ambient component, diffuse). The primary and ambient components are rendered separately to preserve their spatial characteristics, in accordance with the actual playback configurations. The core problem is how to extract the primary and ambient components from channel-based audio content efficiently. To answer this question, this chapter begins with the fundamentals of spatial hearing and reviews existing spatial audio reproduction techniques, as well as prior arts in primary ambient extraction, which is also compared with another sound scene decomposition technique: blind source separation.

Keywords Spatial audio · Fundamentals · Sound scene decomposition · Primary ambient extraction (PAE) · Blind source separation

Spatial audio, also known as three-dimensional (3D) audio, refers to the perception of sound in 3D space and anything that is related to such a perception, including sound acquisition, production, mastering, processing, reproduction, and evaluation of the sound. This book describes the reproduction of 3D sound based on the formats of the audio content. For this purpose, we first review the fundamental principles of human's spatial hearing and discuss various conventional as well as advanced techniques for spatial audio reproduction. After that, a summary of the prior work on primary ambient extraction is presented.

2.1 Basics of Spatial Hearing

With the ears positioned on both sides of our head, humans are capable to perceive sound around us. The perceived sound can be processed by our brain to interpret the meaning of the sound. Equally amazing is our ability to localize sound in the 3D

J. He, *Spatial Audio Reproduction with Primary Ambient Extraction*,
SpringerBriefs in Signal Processing, DOI 10.1007/978-981-10-1551-9_2

space. This capability of localizing sound in 3D space is often referred to as spatial hearing. In this section, we will review the fundamentals of spatial hearing.

2.1.1 How Do We Hear Sound

From a physical point of view, sound waves, emanating from a vibration process (a.k.a., sound source), travel through the air all the way into our ears. Human ears can be broadly separated into three parts: the outer ear, middle ear, and inner ear, as shown in Fig. 2.1 [WHO06]. The pinna of the outer ear picks up the sound and passes through the ear canal to the eardrum of the middle ear. The sound vibrations captured by the eardrum are transformed into nerve signals by the cochlea. These nerve signals travel through the auditory nerve and reach our brain. Our brain can then interpret the sound we hear. Impairment to any parts of the ear would affect our hearing.

2.1.2 How Do We Localize Sound

For a particular sound source in a 3D space, localization of this sound source would involve three dimensions. Clearly, taking the listener (more specifically the head of the listener) as the center of the space, a polar coordinate system is considered to be more appropriate to describe the 3D space. Hence, we describe the three dimensions as distance, azimuth, and elevation, as shown in Fig. 2.2. Distance is the length of the direct line path between the sound source and the center of the head. Horizontal plane refers to the plane that is horizontal to the ground at ear-level height. Median plane is a vertical plane that is perpendicular to the horizontal plane with the same origin at the center of the head. Azimuth θ refers to the angle between the median

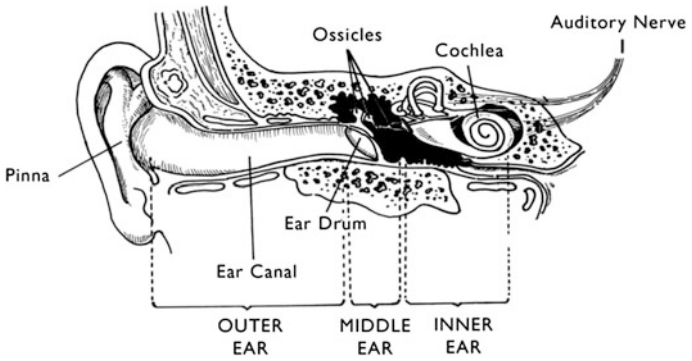

Fig. 2.1 Structure of the human ear (extracted from [WHO06])

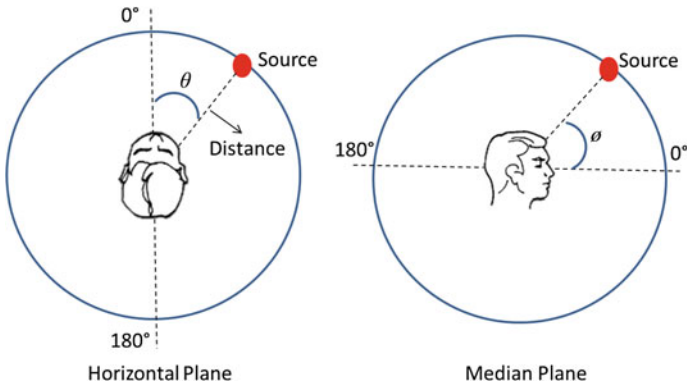

Fig. 2.2 The coordinate system for sound localization

plane and the vector from the center of the head to the source position. Azimuth is usually defined in the clockwise direction, with 0° azimuth referring to the direction right in front of us. Elevation ϕ is defined as the angle between the horizontal plane and the vector from the center of the head to the source position. An elevation of 0° refers to a sound directly in front, and increasing elevation will first move the sound up, then behind, and finally under the listener.

In spatial hearing, sound localization can be considered in different perspectives. In terms of the position of the sound source, we usually consider the direction (i.e., azimuth and elevation) and distance of the sound. Perception of single sound source is different from multiple sound sources, where incoherent sound sources are perceived as separate auditory events and coherent sound sources are governed by summing localization (usually for sound sources with time difference under 1 ms) or precedent effect (for time difference above 1 ms, e.g., reflections) [Bla97]. Coherent sound sources that arrive after several milliseconds would be perceived as echo, which is quite common for sound in enclosed space. For sound localization task, human brains combine various cues from perceived sound and other sensory information such as visual images. It has been commonly known that the following cues contribute to sound localization [Bla97, Beg00, AlD11, Xie13]:

(1) Inter-aural time difference (ITD),
(2) Inter-aural level difference (ILD),
(3) Spectral cues (monaural, relevant to the anthropometry of the listener),
(4) Head movement cues (a.k.a., dynamic cues),
(5) Intensity, loudness cues,
(6) Familiarity to sound source,
(7) Direct-to-reverberation ratio (DRR),
(8) Visual and other non-auditory cues.

Among the seven auditory cues (1) to (7), the first four contribute to direction localization, whereas the last three affect distance perception.

2.1.3 Direction Perception: Azimuth and Elevation

A variety of psychoacoustic experiments have demonstrated human's ability to localize the direction of the sound source. The minimum audible angle (MAA) can reach as low as 1°–3° for broadband sound (e.g., white noise) in the front horizontal plane (±90° azimuth), though it becomes worse for other directions and narrow-band sound [Bla97]. The ITD and ILD are the two most important cues for azimuth direction localization. The ITD refers to the difference of time that the sound travels from the source to the left and right ears. Apparently, sound from different directions would have different traveling time durations to the two ears, resulting different ITDs. The ILD is mainly caused by the attenuation of the sound levels in the contralateral ear (further to the source) due to the head shadowing effect, compared to the ipsilateral ear (nearer to the source). According to the duplex theory [Ray07], ITD relates to the ability of human auditory system to detect inter-aural phase differences at low frequency, and hence, ITD is more dominant in low frequency, whereas ILD dominants at high-frequency region. The cutoff frequency is determined by the distance between the two ears (typically 22–23 cm), which is usually considered to be around 1500 Hz.

For localization of sound in different elevations, ITD and ILD are not enough. This is because identical ITD and ILD values can be obtained from the sound source in a conical surface, as shown in Fig. 2.3 [Beg00]. This is the so-called cone of confusion phenomenon [Mil72]. One of the most common perceptual errors in cone of confusions is the front–back confusions, where one perceives a front (or back) sound in the back (or front). In order to perceive the elevation directions correctly, spectral cues are required. Spectral cues are mainly caused by head, torso, and pinna that filter the incoming sound waves. Sound from different elevations would reach different parts of our body (especially the pinna) and undergoes different reflections before entering the ear canal. Most of the spectral cues due to pinna occur at frequencies above 3 kHz, and the spectral cues due to head and torso appear in lower frequencies. It is worth mentioning that the spectral cues vary greatly from person to person due to the idiosyncratic anthropometry of the listener. In addition to the static cues mentioned above, dynamic cues due to head movement are extremely useful in resolving localization errors, especially front–back confusions.

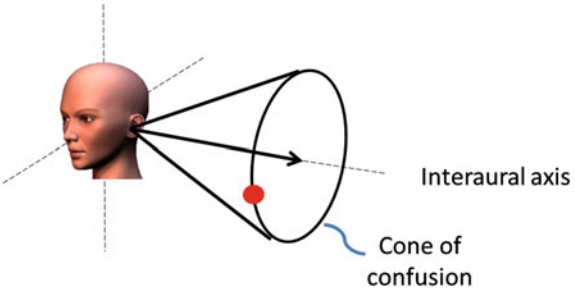

Fig. 2.3 Cone of confusion due to identical ITD and ILD

The head-related transfer function (HRTF) is usually introduced to describe the change in the sound spectra due to the interactions of the sound wave with the listener's head, torso, and pinna, which is defined as follows. In a free-field environment, take the Fourier transform of the sound pressure (SP_L or SP_R) at the eardrums of the two ears and the sound pressure (SP_0) at the center of the head with the listener absent. The HRTF is the ratio of these two Fourier representations. Since human has two ears, HRTF typically comes in pairs. Clearly, HRTF is a function of frequency (f), direction (θ, ϕ), distance (r), and listener (lsn) and is expressed as

$$\begin{aligned} H_L(f,\theta,\phi,r,\mathrm{lsn}) &= \frac{\mathrm{SP}_L(f,\theta,\phi,r,\mathrm{lsn})}{SP_0(f,r)}, \\ H_R(f,\theta,\phi,r,\mathrm{lsn}) &= \frac{\mathrm{SP}_R(f,\theta,\phi,r,\mathrm{lsn})}{SP_0(f,r)}. \end{aligned} \tag{2.1}$$

where P_L, P_R, and P_0 are sound pressures in the frequency domain. According to Algazi et al. [ADM01, BrD98, ADD02], HRTF can be approximated by a structural composite of pinna-less head and torso, and the pinna, which is mainly effective at modifying the source spectra at low and high frequencies, respectively. In the far field, HRTF is usually considered to be independent of distance [Ken95a]. The time-domain representation of HRTF is referred to as head-related impulse response (HRIR).

In Figs. 2.4, 2.5, and 2.6, the HRIR and HRTF of subjects from the CIPIC HRTF database are plotted [ADT01]. The HRIR and HRTF of the same subject at different directions are shown in Fig. 2.4. It is clear that the waveform and magnitude spectra shapes vary with the direction horizontally and vertically. In Fig. 2.5, we show the ITD and ILD (full-band) that are computed from the HRTFs of the same subject. It is clear that ITD and ILD exhibit a close-to-linear relationship with the azimuth, and the change across different elevations is minimal, especially at non-lateral azimuthal directions. The HRIR and HRTF of three different subjects are plotted in Fig. 2.6, which indicates that HRTF generally differs from individual to individual, especially the spectral notches in the high-frequency range. The individual differences of HRTF among different subjects are indeed due to the anthropometric features of these subjects.

2.1.4 Distance Perception

Perception of distance of sound sources is important in sound localization. In sound rendering, it is critical to recreate the perception of distance of the sources close to natural listening. However, the challenges in simulating accurate distance perception are numerous. Human beings' ability to accurately estimate the distance of a sound source has long been known to be poorer compared to our ability to estimate

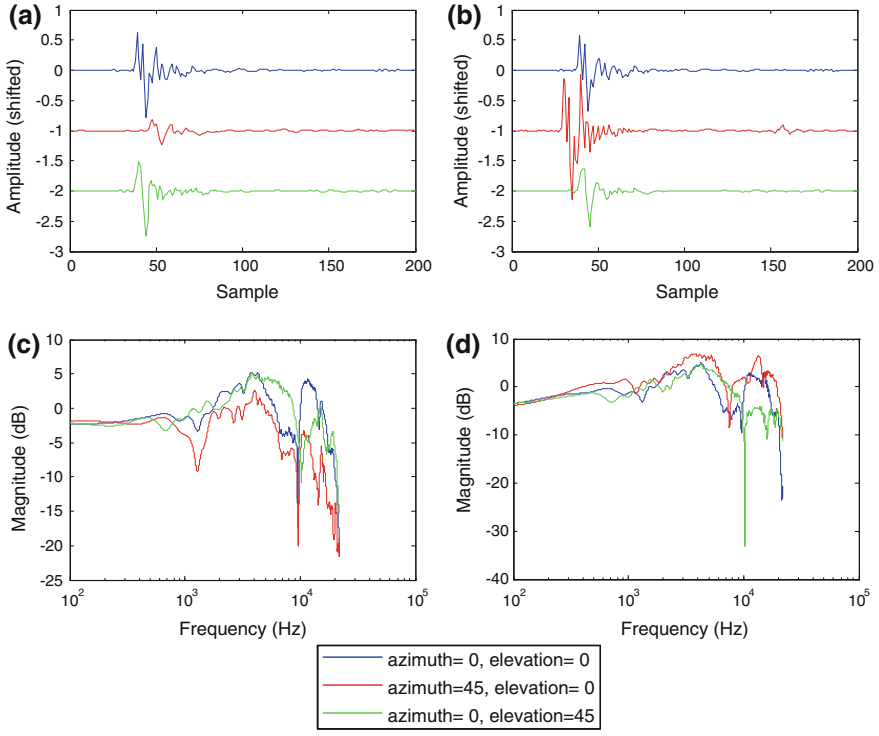

Fig. 2.4 HRIR and HRTF of the same subject (CIPIC HRTF database subject 003 [ADT01]) in different directions

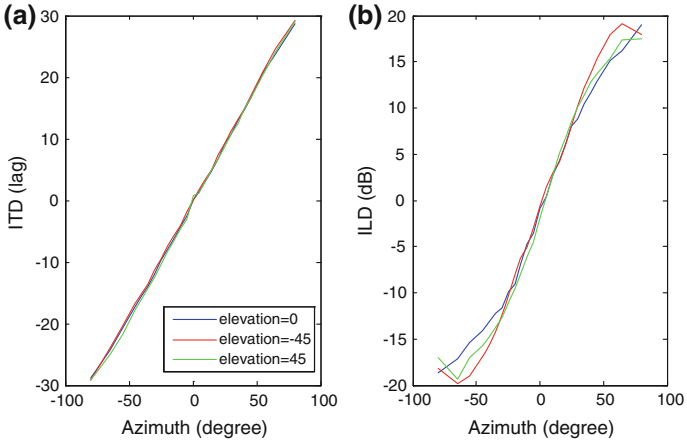

Fig. 2.5 ITD and ILD of the same subject (CIPIC HRTF database subject 003 [ADT01]) in different directions

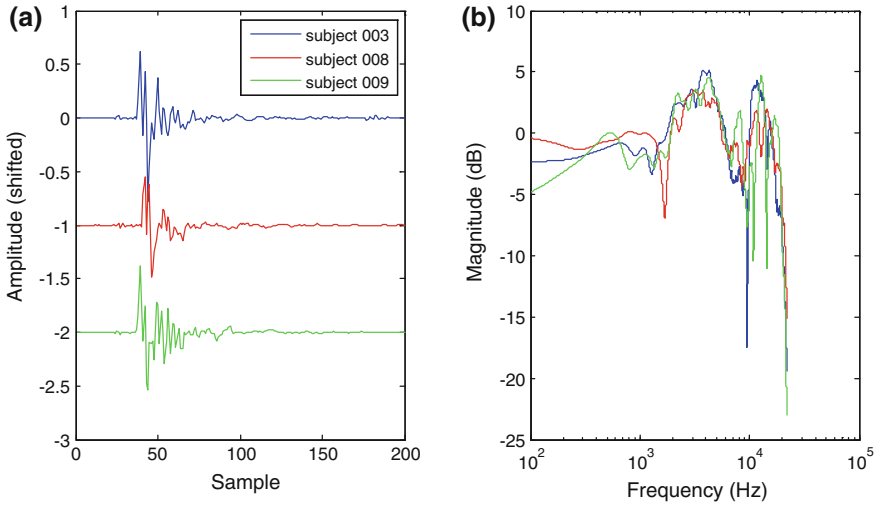

Fig. 2.6 HRIR and HRTF (left ear, azimuth = 0°, elevation = 0°) of three different subjects (subjects 003, 008, 009 in the CIPIC HRTF database [ADT01])

directions, even in the physical listening space [Zah02]. The experiments conducted by Zahorik showed that the perceived distance can usually be expressed in a power function of the actual distance [Zah02a]. The direct-to-reverberation energy ratio is found to be the most critical cue for absolute distance perception, even though the intensity, loudness, and binaural cues (including ILD and inter-aural coherence) can provide relative cues for distance perception [Zar02b, Beg00]. However, accurate simulation of distance perception is challenging since reverberation depends on the room characteristics. The correct amount of reverberation to be added to simulate distance perception in a particular room can be obtained only by carrying out acoustical measurements.

2.1.5 Sound in Rooms: Reflections and Reverberation

Though sound localization is discussed in free-field environment, the real-life sound environment is never free-field. The existing free-field environment can only be found in an anechoic chamber. Rooms that we live in everyday are filled with reflections and reverberation, usually characterized by the room impulse response (RIR). A schematic illustration of RIR is shown in Fig. 2.7. A typical RIR consists of three parts: the direct path, early reflections, and late reverberation (after 80 ms). An important aspect of room acoustics is the reverberation time RT_{60}, which is

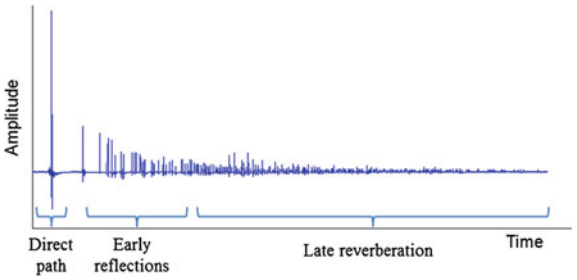

Fig. 2.7 A schematic illustration of RIR (adopted from [VPS12])

defined by the time that it takes for the sound to attenuate by 60 dB once the sound source ceases. To simulate the perception of sound in rooms (or sound environment in general), RIRs that are derived or measured from the (approximately) geometrically identical room are usually used to add artificial reverberation to the dry sound sources [VPS12].

2.1.6 Psychoacoustics and Critical Band

Sound is meaningful when it is perceived by humans. Changes in the physical part of the sound (including frequency, intensity, phase, direction) may not always excite perceptual difference. This is mainly due to the limitation of human auditory system. Thus, in additional to objective evaluation, psychoacoustic experiments, which are in the form of subjective listening tests, are conducted to evaluate the performance of a sound reproduction system. The psychoacoustic experiments could help us better understand how the system actually performs in practice. The psychoacoustic experiments usually include the localization of the sound sources, quality of the synthesized sound, quality of the reproduction system (e.g., loudspeakers and headphones), quality of the rendering methods, and so on.

One of the most important aspects of psychoacoustics is auditory masking, where a louder sound masks (fully or partially) a weaker sound when their spectra are close. Auditory masking happens in frequency domain (spectral masking) and time domain (temporal masking). The range of the spectra for spectral masking is defined based on its critical band, as per the psychoacoustic experiments. According to Zwicker [Zwi61], 24 bands known as the Bark scale are defined to cover the frequency range of human listening. Each critical band has a center frequency with an approximate 1/3 octave bandwidth. The conversion from frequency (f in kHz) into the Bark can be described as:

$$\text{Bark} = 13\arctan(0.76f) + 3.5\arctan\left[(f/7.5)^2\right]. \tag{2.2}$$

Another example of critical band is the equivalent rectangular bandwidth (ERB) [Moo98], which is described as:

$$\mathrm{ERB} = 24.7(4.37f + 1). \tag{2.3}$$

It is widely believed that the human auditory system is performing the critical band analysis of the incoming sound, in tasks like localization and separation of sound [Fle40, Bre90]. Therefore, many audio processing systems are derived based on the concept of critical band (or its equivalents). For example, in binaural cue coding (BCC), 20 non-uniform filterbank based on ERB is employed [FaB03]. Furthermore, MPEG Surround employs a hybrid quadrature mirror filter (QMF) filterbanks [SBP04, HPB05] that match the frequency resolution of the human auditory system.

2.2 Spatial Audio Reproduction

Most of the time, we are not listening to real sound in a real environment, but are listening to a reproduced sound playback from a sound reproduction system. The reproduced sound is often referred to as virtual sound, as compared to real sound in natural listening.

2.2.1 A Brief History of Sound Reproduction Systems

Ever since the invention of phonograph by Edison in 1887, sound has been an essential part of telecommunication and media. The first stereo loudspeaker system was introduced by Blumlein [Blu31] in 1931, which has since then become the most popular sound reproduction system in homes. It takes humans some forty years to come up with new sound systems, including the first Dolby surround sound [ITU12] and Ambisonics invented by Gerzon [Ger73]. Though invented at almost the same time, these two systems undergo extremely different paths. The surround sound reproduction system, including 5.1 and 7.1, as pushed by the film and music industries, has become the most prevalent home theater systems. The 5.1 surround sound system requires five speakers placed at center (0° azimuth), front left (−30° azimuth), front right (30° azimuth), surround left (−110° azimuth), surround right (110° azimuth), and a subwoofer. The 7.1 surround sound system extends 2 surround speakers in 5.1–4 speakers. The multichannel surround sound system keeps evolving, from one layer to two layers (such as 9.1 and 10.2) to even more layers (such as 22.2 [HMS11], Auro 3D). On the other hand, Ambisonics, despite its mathematical beauty (based on Huygens' principle), was not well adopted in commercial systems. Nevertheless, the research on Ambisonics was never stopped in academia and it gains popularity in recent years, as shown in new MPEG-H

standard [HHK14]. In 1993, another sound reproduction technique, wave field synthesis (WFS), was introduced [Ber88, BVV93] and has found its presence in commercial products since 2001. Besides the development of loudspeaker systems, headphones are getting more and more widely used in recent years, which is mainly due to the rapid increase in mobile devices. The HRTFs are widely used in headphone-based 3D sound reproduction [Beg00, AlD11, SHT15]. Today, we see a variety of sound reproduction systems in various applications, from cinema, home theater, to on the go. More and more 3D sound reproduction techniques have been studied and implemented in commercial products.

2.2.2 *Representations of Audio Content*

With the development of different recording and mixing techniques, different types of audio content representations have emerged in commercial market. Three main types are as follows: channel-based, object-based, and transform-domain-based.

Channel-based format has been the most common way of audio content representation. The channel-based format is playback-oriented as the channel signals can be directly fed to the loudspeakers based on the standard configuration (i.e., prescribed positions). Usually, no additional processing (or very little processing like volume control) is required. This is because the channel-based format is usually the outcome of the sound mixing process (performed by the sound engineer). Besides the easy applicability for the playback, the channel-based format is also rather efficient at transmission and storage. The down side of channel-based format lies in its requirement to have a fixed playback system that corresponds to the number of channels. For example, stereo audio content requires the two speakers to be placed symmetrically at $\pm 30°$ azimuth on the two sides of the listener. The 5.1 channel further adds a center and two rear channels, placed at $0°$ and $\pm 110°$ azimuth, respectively, together with a subwoofer (low-frequency effect channel). A matrix system that enables the downward compatibility of 5.1 is discussed in [ITU12]. Other channel-based formats include 7.1, 9.1, 10.2, and all the way up to 22.2 in three vertical layers. Adding height channels in channel-based audio is a fundamental improvement over horizontal loudspeaker setup to make the sound reproduction in full three dimensions. Commercial examples involving height channels on top of the conventional surround sound formats include Dolby ATMOS [Dol13] and Auro 3D [Aur15].

Object-based format is the most original format of a sound recording. Object-based format represents a sound scene using a combination of sound objects with the associated metadata [HHK15]. Sound objects are essentially individual sound sources. The metadata usually consists of two types: static metadata, such as language and on/off time, and dynamic metadata, such as position or direction, level, width, or diffuseness of the sound object. Not all audio objects are separated. Those objects that collectively contribute to a fix sound effect or sound environment shall be grouped and regarded as one “larger” audio object. As a result, metadata

can be specified for each audio object or a group of objects. The greatest benefit of object-based audio is that it can be rendered optimally for any arbitrary playback systems. Meanwhile, interactivity can be enabled, for example, changing to another language of speech, increasing the loudness of certain objects (e.g., speech level shall be higher for hearing-impaired listeners), and adapting the position of the sound objects according to listener's movement in virtual reality applications. The object-based format is the best format in terms of reproduction flexibility and quality. However, two challenges that are found in practical implementation are high storage or transmission bandwidth and high computation complexity for real-time rendering [MMS11]. Important aspects on the implementation of audio objects' coding and rendering were extensively studied in [Pot06]. Some work has been carried out by MPEG to achieve an efficient coding of sound objects based on perceptual features [HPK12].

The other type of audio representation is known as the transform-domain-based format (or scene based, Ambisonics) [SWR13]. Transform-domain-based format encodes the sound scene using orthogonal basis functions physically (using microphones) or digitally. In the reproduction, a corresponding rendering process is required. Though individual sound objects are not used, transform-domain-based format can also achieve flexibility in reproduction for various playback setups, thanks to the sound field analysis and synthesis principle [Pol05]. However, the transform-domain representation is less common and less supported (e.g., recording/reproduction equipment) in industry than in academia.

2.2.3 Spatial Audio Reproduction Techniques

These above-mentioned sound scene representations support different spatial audio reproduction techniques. Due to the nature of channel-based representations, conventional spatial audio reproduction techniques are straightforward as the audio signals of each channel are directly sent to drive the corresponding loudspeaker, resulting in stereo loudspeaker playback, 5.1, 7.1 surround sound playback, and stereo headphone playback. The simplicity of channel-based reproduction is achieved at the cost of strict requirement of exact match of the playback configuration. When there is a mismatch between the audio content and actual playback configuration, the performance is degraded, though simple down-mixing and up-mixing approaches can be applied.

In contrast to the channel-based format, the object-based and transform-domain-based formats are more flexible in the playback and usually achieve better performance in spatial audio reproduction. Modern spatial audio reproduction techniques can usually be divided into two classes, namely the physical reconstruction and perceptual reconstruction [HWZ14].

The first class of physical reconstruction aims at synthesizing the sound field in the listening area or point to be (approximately) equal to the desired sound field. Sound field synthesis is essentially based on the physical principle of synthesizing

acoustic pressure using a weighted distribution of monopole sources [SWR13]. Two examples of sound field synthesis techniques are Ambisonics (4 channels) or high-order Ambisonics (HOA, consists of more than 4 channels), and WFS. Ambisonics or HOA decomposes (or encodes) a sound field using spherical harmonics, which results in the transform-domain-based representation. With more channels, HOA can improve the spatial quality of reproduced sound field over Ambisonics. The best listening area in Ambisonics is usually limited to the central area of the sphere. In contrast, WFS can extend the sweet spot to a much wider area by approximating the propagation of the primary source using an array of secondary sources (loudspeakers). The loudspeaker driving signals are derived using a synthesis system function and source signals, which are expressed in object-based format. Compared to Ambisonics, WFS is not only well studied in academia, but also employed in some commercial sound systems such as IOSONO [Ios15] and Sonic Emotion [SoE15]. A major challenge in the physical reconstruction techniques is the requirement of large amount of loudspeakers and high computational complexity (especially in real-time rendering scenarios) [SWR13].

The other type of spatial audio reproduction techniques is based on the perceptual characteristics of human auditory system that our listening is not very sensitive. A good spatial audio reproduction is one that sounds good. The key idea of perceptual based spatial audio reproduction techniques is to have the sound captured by the listener's eardrum to be perceptually close to the desired sound field. While the reproduced sound field does not always well match the desired sound field, perceptual based spatial audio reproduction techniques can greatly simplify the reproduction method. The simplest example of this category is the amplitude panning techniques, which are widely employed in sound mixing for stereo and surround sound [Hol08]. Techniques that extend amplitude panning to 3D space include the vector base amplitude panning [Pul97, PuK08] and variants like distance-based amplitude panning [LBH09]. Amplitude panning techniques are based on the ILD cues to recreate the correct direction of the sound sources. Similarly, time delay techniques that vary the ITD can also be used for spatial audio reproduction [SWR13].

However, the amplitude panning and time delay techniques are usually too simple to reproduce the correct impression of the sound sources with increased source width [MWC99], degraded location performance [ThP77], and coloration [PKV99]. A better approach is to consider the complete localization cues, which are included in the HRTFs [Beg00]. This approach is usually applied in headphone playback, and it is known as binaural rendering. The key idea in binaural rendering is to consider the sound source propagation process (from sound source to listener's eardrum) as a linear-time-invariant system and express this alteration of the source spectra due to human body as a filter. Therefore, the perception of any source from any direction can be recreated by convolving the sound source with the corresponding filters to obtain the driving signals that are sent to a compensated headphone (assumed transparent). The same concept of binaural rendering can also be applied in stereo loudspeakers, which is known as transaural rendering [Gar97]. Compared to binaural rendering, transaural rendering requires one additional

process known as crosstalk cancelation. Multichannel extension of crosstalk cancelation and transaural system are discussed in [Gar00]. Crosstalk cancelation techniques are very sensitive of listener movement and small changes in sound environment, which limits the practical use of transaural systems. In contrast, binaural rendering over headphones is much more widely used. However, there remain two major challenges. The first one lies in the large variations of the HRTFs among different individuals. The use of non-individualized HRTF will degrade the localization accuracy. The other problem is the headphone itself, which is hardly completely transparent, and headphone effect compensation varies not only from person to person, but also even after repositioning. Due to the advent of virtual/augmented reality applications, many studies on HRTF individualization and headphone compensation are currently being carried out.

2.2.4 Spatial Audio Processing

In order to achieve the goal of efficient, flexible, and immersive spatial audio reproduction, different spatial audio signal processing techniques are introduced. The aim of spatial audio processing (or coding) techniques is to complement the discrepancies in the above-mentioned spatial audio reproduction techniques, with a focus on channel-based signals and conventional multichannel reproduction techniques. Generally, these techniques are based on the concept of parametric spatial audio processing [KTT15] and exploit the perceptual characteristics of human auditory systems [Bla97]. In this part, we focus on five most widely studied frameworks, though other variations could also been found in the literature. Among the five frameworks discussed below, two of them deal with channel-based signals, one on the object-based signals and another one on the transform-domain-based signals, and the latest one consolidates all three types of signals.

For channel-based signals, the objective of spatial audio processing is to achieve a more efficient representation that can reproduce perceptually plausible sound scenes. The most widely known framework comes from the MPEG audio group, known as MPEG Surround [HKB08, BrF07, HiD09]. In MPEG Surround, the multichannel signals go through a spatial analysis process and is represented using a down-mixed version together with the spatial parameters, as shown in Fig. 2.8. In the spatial synthesis, the original multichannel can be reconstructed using the spatial parameters in a way that the spatial perception is maximally preserved. Furthermore, other types of synthesized output include the direct down-mix for the playback with a reduced number of loudspeakers and binaural signals for headphone playback [BrF07, FaB03]. The details on the coding of the spatial parameters can be found in BCC framework [FaB03, BaF03].

Another framework that is also targeting channel-based audio is the so-called spatial audio scene coding (SASC) framework developed by Jot et al. [GoJ08, JMG07, GoJ07a, GoJ06a, GoJ06b, GoJ07c]. Compared to MPEG Surround, SASC was designed to address the pressing need to enhance sound reproduction over

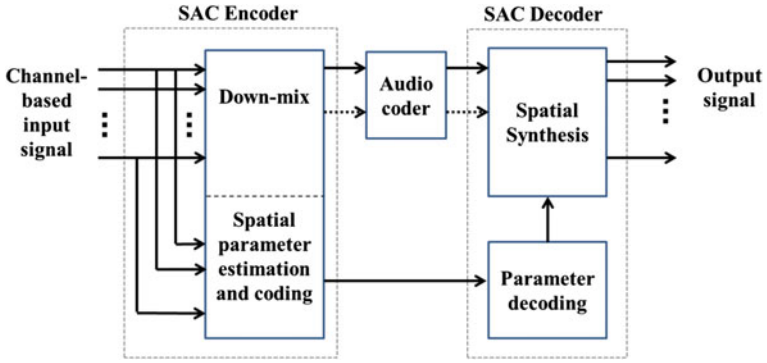

Fig. 2.8 Basic concept of MPEG Surround (adapted from [HiD09])

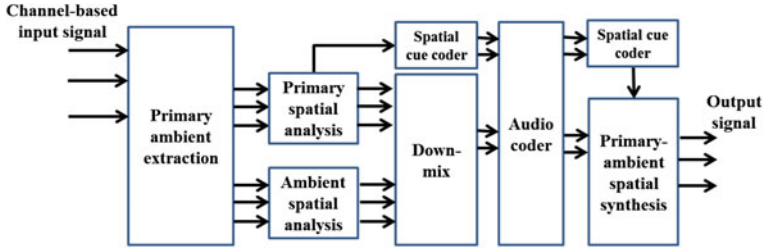

Fig. 2.9 Block diagram of spatial audio scene coding (*SASC*) (adapted from [GoJ08])

arbitrary playback configurations in loudspeakers and headphones. The detailed block diagram is shown in Fig. 2.9. In SASC, a sound scene is considered as a sum of primary and ambient components. Therefore, primary ambient extraction (or decomposition) is applied first, followed by the spatial analysis carried out independently for the primary and ambient components to obtain the spatial cues (i.e., localization information). In the spatial synthesis, the output is reconstructed using the primary and ambient components as well as the spatial cues. By taking into account the actual playback format, the reconstruction is able to fit any playback configuration. Due to this advantage of SASC, the primary ambient extraction work described in this book is essentially based on SASC. Details on the primary ambient extraction will be discussed throughout this book.

For object-based audio signals, MPEG introduced MPEG spatial audio object coding (SAOC) framework in 2012 [HPK12]. Similar to MPEG Surround, the MPEG SAOC aims to achieve an efficient representation of the object-based audio using a parametric approach that takes a down-mix of the audio objects in sub-band with supplementary inter-object information, as shown in Fig. 2.10. In the synthesis, the object decoder can be employed first before the render or can be combined into one block. Based on the information of the actual playback information,

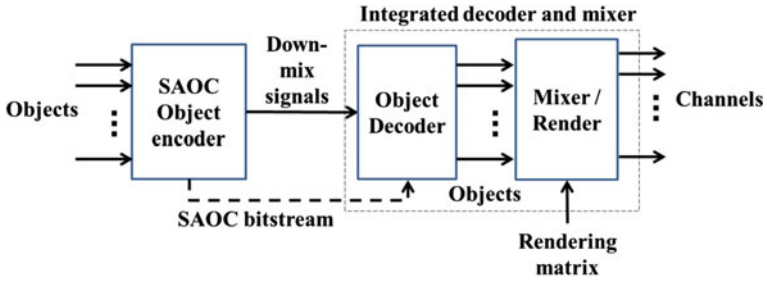

Fig. 2.10 Basic concept of spatial audio object coding (*SAOC*) (adapted from [HPK12])

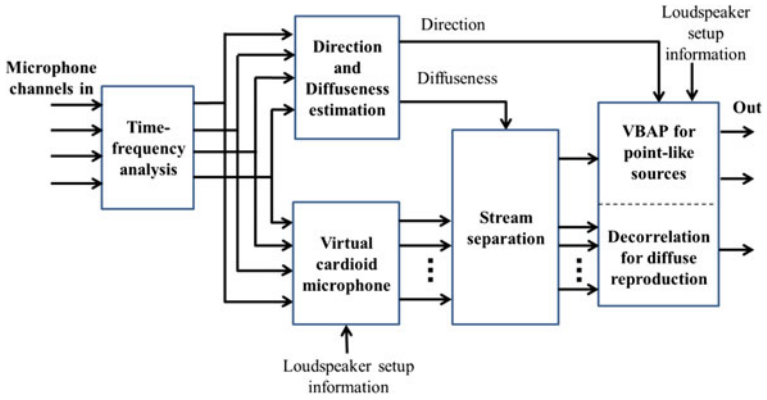

Fig. 2.11 Block diagram of directional audio coding (*DirAC*) (adapted from [Pul07])

a rendering matrix is used to transform the audio objects into channel signals for playback. It shall be noted that SAOC can also achieve the flexibility and interactivity of the object-based format.

For transform-based signals, a parametric spatial audio processing framework known as directional audio coding (DirAC) was introduced by Pulkki et al. [Pul07]. As shown in Fig. 2.11, DirAC analyzes the direction and diffuseness information of the microphone signals (in B-format) and then decomposes the microphone signals into two streams, namely diffuse streams and non-diffuse streams. As shown in Fig. 2.11, these two streams go through different rendering processes, where the non-diffuse streams are processed using VBAP with the loudspeaker setup information provided, and diffuse streams are decorrelated and played back over all the channels. The advantage of such decomposition, similar to SASC, is to be able to achieve flexible reproduction over arbitrary playback configurations.

Finally, MPEG-H [HHK14, HHK15], introduced in 2014, aims to handle all three types of audio content (channel-based, object-based, and transform-domain-based, presenting a complete solution for universal spatial audio reproduction. An overview of MPEG-H framework is depicted in Fig. 2.12. In the first step, the input

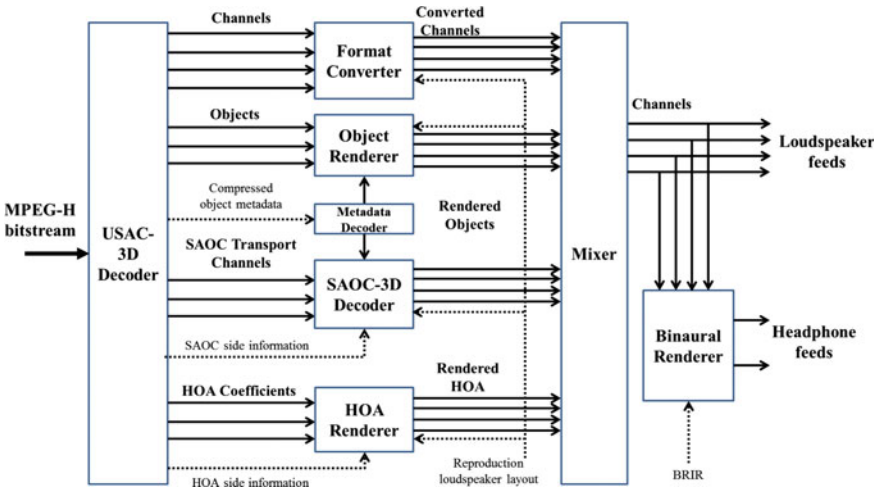

Fig. 2.12 Overview of MPEG-H 3D audio coding (adapted from [HHK14])

bit stream is converted to their respective format using unified speech and audio coding (USAC)-3D core decoder. Next, different content types go through corresponding processing before they were mixed into channel signals that match the actual playback system layout. Finally, in the case of headphone playback, a binaural rendering of loudspeaker signals based on binaural room impulse response (BRIR) is employed. With such a unified framework, MPEG-H 3D audio can be employed for any content type and any playback configuration, while achieving the highest spatial audio quality.

2.2.5 Spatial Audio Evaluation

In spatial audio reproduction, the quality of the reproduced sound scene is usually evaluated on human perception. Perceptual evaluation of audio quality is often achieved using subjective listening tests [BeZ07]. Unlike conventional sound quality evaluation that usually only considers the timbre quality [GaS79] (e.g., evaluation of the quality of audio codec [ITU03]), the spatial quality is equally important in spatial audio evaluation [Rum02]. Referring to these two aspects of audio quality for spatial audio evaluation, Table 2.1 below summarizes the various attributes that can be considered in each category [SWR13]. Among the timbre attributes, timbre fidelity, coloration, and distortion are more widely used. For spatial attributes, spatial fidelity, envelopment, distance, and localization are more important. Relative importance between the spatial quality and timbre quality is investigated in [RZK05], and it was summarized that the overall sound quality can be explained by the sum of 70 % of the timbre quality and 30 % of the spatial quality. Beyond these "perceptive domain" attributes as listed in Table 2.1, the

Table 2.1 Attributes used for perceptual spatial audio evaluation (adapted from [SWR13])

Category	Attribute	Description
Timbre	Timbral fidelity	Degree to which timbral attributes agree with reference
	Coloration	Timbre change considered as degradation of auditory event
	Timbre, color of tone	Timbre of auditory events
	Volume, richness	Perceived thickness
	Brightness	Perceived brightness or darkness
	Clarity	Absence of distortion, clean sound
	Distortion, artifacts	Noise or other disturbances in auditory event
Spatial	Spatial fidelity	Degree to which spatial attributes agree with the reference
	Spaciousness	Perceived size of environment
	Width	Individual or apparent source width
	Ensemble width	Width of the set of sources present in the scene
	Envelopment	Degree to which the auditory scene is enveloping the listener
	Distance	Sense of perspective in the auditory scene as a whole
	Externalization	Degree to which the auditory event is localized inside or outside of the head
	Localization	Measure of how well a spatial location can be attributed to an auditory event
	Robustness	Degree to which the position of an auditory event changes with listener movements
	Stability	Degree to which the location of an auditory event changes over time

highest level of perception is in the "affective domain" [11], where the listeners indicate their preference of the perceived sound scenes. In spatial audio reproduction where virtual audio is presented to the listener, an importance affective feature is the immersiveness. In other words, while listening to the reproduced sound, how much the listener feels as if him/herself is inside the virtual scene (a.k.a., being there). Pursuing an immersive reproduction is the common aim of all spatial audio reproduction systems including the primary ambient extraction-based spatial audio reproduction.

2.2.6 Summary and Comparison of Spatial Audio Reproduction

Table 2.2 summarizes the advantages, disadvantages, and the status of the three audio formats discussed in this section. Furthermore, the spatial audio reproduction systems that correspond to each audio format are listed, together with the possible

Table 2.2 A summary of the characteristics of three audio content formats and their relationships with the spatial audio reproduction systems and processing techniques

Audio content format	Channel-based	Object-based	Transform-domain-based
Advantages	Easy to set up; no processing for the matched playback configurations	Flexible for arbitrary playback configuration; accurate sound image; enable interactivity	Flexible for arbitrary playback configuration; full 3D sound image
Disadvantages	Difficult to fit in different playback configurations; 3D sound image limited	High transmission or storage; high computation complexity	Require a large number of speakers placed on the surface of a sphere
Status	Legacy audio format, still dominant	Emerging audio format; gaining popularity	Not well adopted commercially
Desired reproduction system	Stereo and multichannel surround sound system	Amplitude panning, WFS, binaural, transaural rendering	Ambisonics and HOA
Typical spatial audio processing	MPEG Surround [HiD09], SASC [GoJ08]	SAOC [HPK12]	DirAC [Pul07]

spatial audio processing techniques. It shall be noted that though classified in Table 2.2, there are still exceptions that link one audio format with other reproduction systems or processing techniques. For example, channel-based signals can also be employed in binaural/transaural rendering by considering one channel as one audio object with a fixed position. Ambisonics reproduction can also be extended to object-based audio by encoding the sound objects using spherical harmonics. It could be foreseen that with the advancement of semiconductor industry, the efficiency problem in object-based audio could be greatly alleviated and object-based audio will overtake channel-based to become the most commonly used audio format. Thus, advanced spatial audio reproduction system can essentially be employed in homes and mobile platforms. Nevertheless, there is still a need to ensure the compatible playback of channel-based audio signals due to the large amount of content available today.

2.3 Prior Work in Primary Ambient Extraction

In this section, we will summarize various existing works on PAE and highlight how our works differ from those in the literature.

As discussed above, PAE is an integral part of SASC framework that considers the audio scene as a sum of the primary components and ambient components. The

primary components are usually composed of directional pointlike sources, whereas the ambient components are diffuse sound determined by the sound environment. The target audio format of PAE is channel-based signals. Therefore, we classify the PAE approaches based on the number of channels in the input signals: single channel (or mono), stereo, and multichannel. From another perspective, the complexity of the audio scenes affects the performance of PAE greatly. Based on the existing PAE work, the complexity of audio scenes can generally be classified into three levels, namely basic, medium, and complex. The basic complexity level refers to the audio scene where there is usually one dominant source in the primary components, with its direction created using only amplitude panning techniques. More specific conditions for the basic level will be detailed in Chap. 3. The medium complexity level requires only the condition of one dominant sources, without restricting how its direction (using amplitude panning, delay, or HRTF, etc.) can be created. In the complex audio scene level, we consider multiple dominant sources in the primary components. The number of dominant sources in this case is also usually limited to 2–3 since it is impractical for listeners to concentrate on too many sources at one time and listeners would rather consider those sources as ambient components. Note that those PAE approaches that claimed to work in multiple sources using sub-band techniques, but without detailed study, will not be classified in the complex level category. From these two perspectives, we shall classify the existing PAE approaches into different categories, as summarized in Table 2.3.

With a glance of this table, it is observed that most of the PAE works are mainly focused on the stereo signals, due to the large amount of stereo content. There are some works carried out for multichannel signals, whereas very limited works are on single-channel signals. This makes sense because dealing with multichannel signals is much less challenging than dealing with single-channel signals, where there is very limited information (especially the inter-channel relations). Next, we will summarize the PAE work in each category.

2.3.1 Stereo Signals

PAE for stereo signals in the basic complexity category can be classified into four types: (i) time–frequency masking, (ii) principal component analysis (PCA), (iii) least squares (LS), and (iv) ambient spectrum estimation, as well as some other techniques.

One of the earliest works in primary or ambient extraction was from Avendano and Jot in 2002 [AvJ02]. In this work, a time–frequency masking approach was proposed to extract ambient components $\hat{A}_c$ from stereo signals X_c, as

$$\hat{A}_c(m,l) = X_c(m,l)\Psi_A(m,l), \tag{2.4}$$

Table 2.3 An overview of recent work in PAE

No. of channels	Complexity of audio scenes		
	Basic (single source, only amplitude panning)	Medium (single source)	Complex (multiple sources)
Stereo	Time–frequency masking: [AvJ02], [AvJ04], [MGJ07], [Pul07] PCA: [IrA02], [BVM06], [MGJ07], [GoJ07b], [BaS07], [God08], [JHS10], [BJP12], [TaG12], [TGC12], [LBP14] Least squares: [Fal06], [Fal07], [JPL10], [FaB11], [UhH15] Linear estimation: [HTG14] Ambient spectrum estimation: [HGT15a], [HGT15b] Others: [BrS08], [MeF10], [Har11]	LMS: [UsB07] Shifted PCA: [HTG13] Time-shifting: [HGT15c]	PCA: [DHT12], [HGT14], [HeG15],
Multichannel	PCA: [GoJ07b] Others: [GoJ07a], [WaF11], [TGC12], [CCK14]	ICA and time–frequency masking: [SAM06] Pairwise correlations: [TSW12] Others: [StM15]	ICA: [HKO04]
Single	NMF: [UWH07] Neural network: [UhP08]		

Notes
1. Those papers that do not explicitly study and evaluate complex signals will be classified into the basic or medium complexity categories
2. Blue color represents application papers, where no detailed study is carried out on PAE
3. Red color represents our works, which are described in the following chapters of this book

where c denotes the channel index and $0 \leq \Psi_A(m,l) \leq 1$ is the real-valued ambient mask at time–frequency bin (m,l). The time–frequency regions that present high coherence correspond to stronger primary components, and low-coherence time–frequency regions can be attributed to stronger ambient components [AvJ04]. Thus, they derived the ambient mask using a nonlinear function of the inter-channel coherence. The following works on time–frequency masking derive the ambient mask based on the characteristic that ambient components have equal level in the two channels of the stereo signal [MGJ07] or using diffuseness measured from B-format microphone recordings [Pul07].

PCA has been the most widely studied PAE approach [IrA02, BVM06, MGJ07, GoJ07b, BaS07, God08, JHS10, BJP12, TaG12, TGC12, LBP14, HTG14]. The key

idea behind the PCA-based PAE approach is to extract the principal component with the largest variance as the primary components (as the name suggests). Variants of PCA include the modified PCA that ensures uncorrelated ambience extraction [God08], enhanced post-scaling to restore the correct primary-to-ambient energy ratio [JHS10], and correct power of primary and ambient components [BJP12]. In our work [BJP12], we derived a simplified solution for PCA and conducted a comprehensive objective evaluation of PCA, which leads us to the applications of PCA in PAE.

Least-squares algorithm is another type of commonly used PAE approaches [Fal06, Fal07, JPL10, FaB11, HTG14, UhH15]. Based on the basic stereo signal model, least-squares algorithm derives the estimated primary and ambient components by minimizing the mean square error (MSE) of the estimation of these components [Fal06]. Several variants of LS have been proposed and studied in our work [HTG14]. Combining PCA with LS, we proposed a unified linear estimation framework for PAE [HTG14], where details of liner estimation-based PAE can be found in Chap. 3. Furthermore, other least-squares variants were introduced to improve the spatial quality of the extracted primary and ambient components [JPL10, UhH15].

To solve the problem of removing uncorrelated (undesired) ambient components from the extraction output, a new framework based on ambient spectrum estimation was introduced recently [HGT15a, HGT15b]. Details on the ambient spectrum estimation approaches can be found in Chap. 4 of this book. Other PAE approaches that fall into this category include [BrS08] that derives an out-of-phase signal as ambient components; [MeF10] that considers ambient components as the sum of a common component and an independent component; and [Har11] that classifies various signal models for respective extraction.

In order to handle stereo signals that consist of primary components whose directions are created using time/phase differences (i.e., medium complexity), several works can be found in the literature. Usher and Benesty proposed an adaptive approach using normalized least mean squares (NLMS) to extract reverberation from stereo microphone recordings [UsB07]. However, this adaptive approach cannot always yield a good performance in a short time. In contrast, our proposed shifted PCA [HTG13] and extended time-shifting technique [HGT15c] is much simpler in solving this problem. Details on this approach can be found in Chap. 5 of this book.

With respect to stereo signals with multiple sources, there is less work reported in the literature of PAE. One prior work by Dong et al. applied PCA in polar coordinates to reduce the coding noise of stereo signals for multiple source cases [DHT12]. However, the extraction performance was not studied. To fill this gap, we conducted two works that studied PCA with different frequency partitioning methods in frequency domain [HGT14] and PCA with multiple time shifts in time domain [HeG15]. Details are described in Chap. 6 of this book.

2.3.2 Multichannel Signals

Besides the extensive study on PAE for stereo signals, PAE on multichannel signals is less well studied. PCA was originally proposed to work for multichannel signals with only one dominant amplitude-panned source in [GoJ07b]. There are several works [GoJ07a, WaF11, TGC12, CCK14] that only briefly mention the idea for multichannel PAE without in-depth studies. For other multichannel signals with one dominant source, independent component analysis (ICA) can be combined with time–frequency masking to extract the dominant sources [Sam06]. Another approach that was extended from [AvJ04] achieves primary ambient extraction using a system of pairwise correlation. Recently, Stefanakis introduced W-disjoint orthogonality (WDO) and PCA-based foreground suppression techniques in multichannel microphone recordings [StM15]. In the case of multiple sources in multichannel signals, blind source separation techniques can be employed for the purpose of primary ambient extraction. When the number of dominant sources is equal to or less than the number of channels (as it is the case for PAE), ICA is a common technique [HKO04]. Compared to stereo signals, PAE with multichannel signals is in fact easier to solve since there are more data available. Moreover, PAE approaches based on stereo signals can be extended to multichannel signals. Some discussions on this topic can be found in [HeG15b].

2.3.3 Single-Channel Signals

In contrast to stereo and multichannel signals, PAE with single-channel signals is quite challenging due to the limited amount of information available. A critical problem in the single-channel case is that how primary and ambient components can be defined and characterized since there are no inter-channel cues. Nevertheless, two works from Uhle shed some light on solving such a problem. In [UWH07], it is considered that ambient components exhibit a less repetitive and constructive spectra structure than primary components. Therefore, when applying nonnegative matrix factorization (NMF) on the single-channel signal, primary components are better explained and factorized, and the residue can thus be considered as ambient components. However, the NMF method suffers from high computational complexity and latency. To avoid this problem, Uhle and Paul introduced a supervised learning approach for ambient extraction from single-channel signals [UhP08], where a neural network is trained to obtain an ambient spectra mask. Subjective listening tests in [UhP08] validated the improved perceptual quality of the up-mix systems employing these PAE approaches.

2.4 Sound Scene Decomposition Using PAE and BSS

To achieve a flexible and immersive 3D sound rendering, two important constituents of the sound scenes are required, namely the individual sound sources and characteristics of the sound environment. However, this information is usually not directly available to the end user. One has to work with the existing digital media content that is available, i.e., the mastered mix distributed in channel-based formats (e.g., stereo, 5.1). Therefore, it is necessary to extract the sound sources and/or sound environment from their mixtures. Two types of techniques that can be applied in sound scene decomposition are PAE and BSS.

2.4.1 Decomposition Using BSS

Extracting the sound sources from the mixtures, often referred to as BSS, has been extensively studied in the last few decades. In BSS, the sound scene is considered to be the sum of distributed sound sources. The basic mixing model in BSS can be considered as anechoic mixing, where the sources $s_k(n)$ in each mixture $x_c(n)$ have different gains g_{ck} and delays τ_{ck}. Hence, the anechoic mixing is formulated as follows:

$$x_c(n) = \sum_{k=1}^{K} g_{ck} s_k(n - \tau_{ck}) + e_c(n), \ \forall c \in \{1, 2, \ldots, C\}, \tag{2.5}$$

where $e_c(n)$ is the noise in each mixture, which is usually neglected for most cases. Note that estimating the number of sources is quite challenging and it is usually assumed to be known in advance [HKO04]. This formulation can be simplified to represent instantaneous mixing by ignoring the delays or can be extended to reverberant mixing by including multiple paths between each source and mixture. An overview of the typical techniques applied in BSS is listed in Table 2.4.

Based on the statistical independence and non-Gaussianity of the sources, ICA algorithms have been the most widely used techniques in BSS to separate the sources from mixtures in the determined case, where the numbers of mixtures and

Table 2.4 Overview of typical techniques in BSS

Objective: To extract K ($K > 2$) sources from C mixtures		
Case		Typical techniques
Determined: $K = C$		ICA [HKO04]
Overdetermined: $K < C$		ICA with PCA or LS [HKO04]
Underdetermined: $K > C$	$C > 2$	ICA with sparse solutions [HKO04, PBD10]
	$C = 2$	Time–frequency masking [YiR04]
	$C = 1$	NMF [Vir06, VBG14]; CASA [WaB06]

sources are equal [HKO04]. In the overdetermined case, where there are more mixtures than sources, ICA is combined with PCA to reduce the dimension of the mixtures, or combined with LS to minimize the overall MSE [HKO04]. In practice, the underdetermined case is the most common, where there are fewer mixtures than sources. For the underdetermined BSS, sparse representations of the sources are usually employed to increase the likelihood of sources to be disjoint [PBD10]. The most challenging underdetermined BSS is when the number of mixtures is two or lesser, i.e., in stereo and mono signals.

Stereo signals (i.e., $C = 2$), being one of the most widely used audio format, have been the focus in BSS. Many of these BSS techniques can be considered as time–frequency masking and usually assume one dominant source in one time–frequency bin of the stereo signal [YiR04]. In these time–frequency masking-based approaches, a histogram for all possible directions of the sources is constructed, based on the range of the binwise amplitude and phase differences between the two channels. The directions, which appear as peaks in the histogram, are selected as source directions. These selected source directions are then used to classify the time–frequency bins and to construct the mask. For every time–frequency bin (m, l), the kth source at cth channel $\hat{S}_{ck}(n, l)$ is estimated as:

$$\hat{S}_{ck}(m, l) = \Psi_{ck}(m, l) X_c(m, l), \tag{2.6}$$

where the mask and the mth mixture are represented by $\Psi_{ck}(m, l)$ and $X_c(m, l)$, respectively.

In the case of single-channel (or mono) signals, the separation is even more challenging since there is no inter-channel information. Hence, there is a need to look into the inherent physical or perceptual properties of the sound sources. NMF-based approaches are extensively studied and applied in single-channel BSS in recent years. The key idea of NMF is to formulate an atom-based representation of the sound scene [Vir06], where the atoms have repetitive and non-destructive spectral structures. NMF usually expresses the magnitude (or power) spectrogram of the mixture as a product of the atoms and time-varying nonnegative weights in an unsupervised manner. These atoms, after being multiplied with their corresponding weights, can be considered as potential components of sources [VBG14]. Another technique applied in single-channel BSS is the computational auditory scene analysis (CASA) that simulates the segregation and grouping mechanism of human auditory system [WaB06] on the model-based representation (monaural case) of the auditory scenes. An important aspect worth considering is the directions of the extracted sources, which can usually come as a by-product in multichannel BSS. In single-channel BSS, this information of source directions has to be provided separately.

2.4.2 A Comparison Between BSS and PAE

Both BSS and PAE are extensively applied in sound scene decomposition, and a comparison between these approaches is summarized in Table 2.5. The common objective of BSS and PAE is to extract useful information (mainly the sound sources and their directions) about the original sound scene from the mixtures and to use this information to facilitate natural sound rendering. Following this objective, there are three common characteristics in BSS and PAE. First, only the mixtures are available and usually no other prior information is given. Second, the extraction of the specific components from the mixtures is based on certain signal models. Third, both techniques require objective and subjective evaluation.

As discussed earlier, the applications of different signal models in BSS and PAE lead to different techniques. In BSS, the mixtures are considered as the sums of multiple sources, and the independence among the sources is one of the most important characteristics. In contrast, the mixing model in PAE is based on human perception of directional sources (primary components) and diffuse sound environment (ambient components). The perceptual difference between primary and ambient components is due to the directivity of these components that can be characterized by their correlations. The applications that adopted BSS and PAE also

Table 2.5 Comparison between BSS and PAE in sound scene decomposition

	BSS	PAE
Objective	To obtain useful information about the original sound scene from given mixtures and facilitate natural sound rendering	
Common characteristics	• Usually no prior information, only mixtures • Based on certain signal models • Require objective as well as subjective evaluation	
Basic mixing model	Sums of multiple sources (independent, non-Gaussian, etc.)	Primary components (highly correlated) + ambient components (uncorrelated)
Techniques	ICA [HKO04], sparse solutions [PBD10], time–frequency masking [YiR04], NMF [Vir06, VBG14], CASA [WaB06], etc.	PCA [MGJ07], LS [Fal06, HTG14], time–frequency masking [AvJ04, MGJ07], time/phase-shifting [HTG13, HGT14], etc.
Typical applications	Speech, music	Movie, gaming
Related applications	Speech enhancement, noise reduction, speech recognition, music classification	Sound reproduction sound localization coding
Limitations	• Small number of sources • Sparseness/disjoint • No/simple environment	• Small number of sources • Sparseness/disjoint • Low ambient power • Primary ambient components uncorrelated

have distinct differences. BSS is commonly used in speech and music applications, where the clarity of the sources is usually more important than the effect of the environment. On the other hand, PAE is more suited for the reproduction of movie and gaming sound content, where the ambient components also contribute significantly to the naturalness and immersiveness of the sound scenes. Subjective experiments revealed that BSS- and PAE-based headphone rendering can improve the externalization and enlarge the sound stage with minimal coloration [BrS08]. It shall be noted in certain cases, such as extracting sources from their reverberation, BSS shares a similar objective as PAE and hence can be applied in PAE [SRK12].

Despite the recent advances in BSS and PAE, the challenges due to the complexity and uncertainty of the sound scenes still remain to be resolved. One common challenge in both BSS and PAE is the increasing number of audio sources in the sound scenes, while only a limited number of mixtures (i.e., channels) are available. In certain time–frequency representations, the sparse solutions in BSS and PAE would require the sources to be sparse and disjoint [PBD10]. Considering the diversity of audio signals, finding a robust sparse representation for different types of audio signals is extremely difficult. The recorded or post-processed source signals might even be filtered due to physical or equivalently simulated propagation and reflections. Moreover, the audio signals coming from adverse environmental conditions (including reverberation and strong ambient sound) usually degrade the performance of the decomposition. These difficulties can be addressed by studying the features of the resulting signals and by obtaining more prior information on the sources, the sound environment, the mixing process [VBG14], and combining auditory information with visual information of the scene.

2.5 Conclusions

In this chapter, we reviewed the basics on spatial hearing of humans, where the binaural cues are very important. Various aspects on spatial audio reproduction are further discussed, which begins with the history of spatial audio reproduction. Three types of audio representations are explained and found to be deterministic in choosing the appropriate spatial audio reproduction techniques as well as spatial audio processing techniques. With the aim to improve the reproduction flexibility and quality of channel-based audio, primary ambient extraction is introduced. Various PAE approaches are classified and reviewed in this section. The details on the work to improve the performance of PAE in various circumstances will be presented in the following chapters.

References

[AlD11] Algazi VR, Duda RO (2011) Headphone-based spatial sound. IEEE Signal Process Mag **28**(1), 33–42

[ADT01] Algazi VR, Duda RO, Thompson DM, Avendano C (2001a) The CIPIC HRTF database. In: Proceedings IWASPAA, New Paltz, NY, USA

[ADM01] Algazi VR, Duda RO, Morrison RP, Thompson DM (2001b) Structural composition and decomposition of HRTFs. In: Proceedings of IEEE WASSAP, New Paltz, NY

[ADD02] Algazi VR, Duda RO, Duraiswami R, Gumerov NA, Tang Z (2002) Approximating the head-related transfer function using simple geometric models of the head and torso. J Acoust Soc Am 112(5):2053–2064

[AvJ02] Avendano C, Jot JM (2002) Ambience extraction and synthesis from stereo signals for multi-channel audio up-mix. In: Proceedings of ICASSP, pp 13–17

[AvJ04] Avendano C, Jot JM (2004) A frequency-domain approach to multichannel upmix. J Audio Eng Soc 52(7/8):740–749

[Aur15] AURO-3D Concept (2015). Available online: http://www.auro-3d.com/system/concept/. 1 May 2015

[BJP12] Baek YH, Jeon SW, Park YC, Lee S (2012) Efficient primary-ambient decomposition algorithm for audio upmix. In: Proceedings of audio engineering society 133rd convention, San Francisco

[BaS07] Bai MR, Shih GY (2007) Upmixing and downmixing two-channel stereo audio for consumer electronics. IEEE Trans Consum Electron 53(3):1011–1019

[BaF03] Baumgarte F, Faller C (2003) Binaural cue coding-Part I: psychoacoustic fundamentals and design principles. IEEE Trans Speech Audio Process 11(6):509–519

[BeZ07] Bech S, Zacharov N (2007) Perceptual audio evaluation-theory, method and application. Wiley, New York

[Beg00] Begault DR (2000) 3-D sound for virtual reality and multimedia. AP Professional, Cambridge

[BVV93] Berkhout A, de Vries D, Vogel P (1993) Acoustic control by wave field synthesis. J Acoust Soc Am 93(5):2764–2778

[Ber88] Berkhout A (1988) A holographic approach to acoustic control. J Audio Eng Soc 36 (12):977–995

[Bla97] Blauert J (1997) Spatial hearing: the psychophysics of human sound localization. MIT Press, Cambridge

[Blu31] Blumlein AD (1931) Improvements in and relating to sound-transmission, sound-recording and sound reproducing systems. British Patent 394 325

[BrF07] Breebaart J, Faller C (2007) Spatial audio processing: MPEG surround and other applications. Wiley, Hoboken

[BrS08] Breebaart J, Schuijers E (2008) Phantom materialization: a novel method to enhance stereo audio reproduction on headphones. IEEE Trans Audio Speech Lang Process 16(8):1503–1511

[Bre90] Bregman AS (1990) Auditory scene analysis. MIT Press, Cambridge

[BVM06] Briand M, Virette D, Martin N (2006) Parametric representation of multichannel audio based on principal component analysis. In: Proceedings of Audio Engineering Society 120th convention, Paris

[BrD98] Brown CP, Duda RO (1998) A structural model for binaural sound synthesis. IEEE Trans Speech Audio Process 6(5):476–488

[CCK14] Chung H, Chon SB, Kim S (2014) Flexible audio rendering for arbitrary input and output layouts. In: Proceedings of 137th AES convention, Los Angeles, CA

[Dol13] Dolby Atmos-Next Generation Audio for Cinema (White Paper) (2013) Available online: http://www.dolby.com/uploadedFiles/Assets/US/Doc/Professional/Dolby-Atmos-Next-Generation-Audio-for-Cinema.pdf

[DHT12] Dong S, Hu R, Tu W, Zheng X, Jiang J, Wang S (2012) Enhanced principal component using polar coordinate PCA for stereo audio coding. In: Proceedings of ICME, Melbourne, Australia, pp 628–633
[Fal06] Faller C (2006) Multiple-loudspeaker playback of stereo signals. J Audio Eng Soc 54 (11):1051–1064
[Fal07] Faller C (2007) Matrix surround revisited. In: Proceedings of 30th AES international conference, Saariselka, Finland
[FaB03] Faller C, Baumgarte F (2003) Binaural cue coding-Part II: Schemes and applications. IEEE Trans Speech Audio Process 11(6):520-531
[FaB11] Faller C, Breebaart J (2011) Binaural reproduction of stereo signals using upmixing and diffuse rendering. In: Proceedings of 131th audio engineering society convention, New York
[Fle40] Fletcher H (1940) Auditory patterns. Rev Mod Phys 12(1):47–65
[GaS79] Gabrielsson A, Sjogren H (1979) Perceived sound quality of sound reproducing systems. J Acoust Soc Am 65(4):1019–1033
[Gar00] Garas J (2000) Adaptive 3D sound systems. Springer, Berlin
[Gar97] Gardner W (1997) 3-D audio using loudspeakers. PhD thesis, School of Architecture and Planning, MIT, USA
[Ger73] Gerzon MA (1973) Perophony: with-height sound reproduction. J Audio Eng Soc 21(1):3–10
[God08] Goodwin M (2008) Geometric signal decompositions for spatial audio enhancement. In: Proceedings of ICASSP, Las Vegas, pp 409–412
[GoJ06a] Goodwin M, Jot J-M (2006a) A frequency-domain framework for spatial audio coding based on universal spatial cues. In: Proceedings of 120th AES convention
[GoJ06b] Goodwin M, Jot J-M (2006b) Analysis and synthesis for universal spatial audio coding. In: Proceedings of 121st AES convention
[GoJ07a] Goodwin M, Jot JM (2007a) Binaural 3-D audio rendering based on spatial audio scene coding. In: Proceedings of 123rd AES convention, New York
[GoJ07b] Goodwin M, Jot JM (2007b) Primary-ambient signal decomposition and vector-based localization for spatial audio coding and enhancement. In: Proceedings of ICASSP, Hawaii, pp 9–12
[GoJ07c] Goodwin M, Jot JM (2007c) Multichannel surround format conversion and generalized up-mix. In: Proceedings of AES 30th International conference
[GoJ08] Goodwin M, Jot JM (2008) Spatial audio scene coding. In: Proceedings of 125th AES convention, San Francisco, 2008
[HMS11] Hamasaki K, Matsui K, Sawaya I, Okubo H (2011) The 22.2 multichannel sounds and its reproduction at home and personal environment. In: Proceedings of AES 43rd International conference audio for wirelessly networked personal devices, Pohang, Korea
[Har11] Härmä A (2011) Classification of time-frequency regions in stereo audio. J Audio Eng Soc 59(10):707–720
[HeG15a] He J, Gan WS (2015a) Multi-shift principal component analysis based primary component extraction for spatial audio reproduction. In: Proceedings of ICASSP, Brisbane, Australia, pp 350–354
[HeG15b] He J, Gan WS (2015b) Applying primary ambient extraction for immersive spatial audio reproduction. In: Proceedings of APSIPA ASC, Hong Kong
[HTG13] He J, Tan EL, Gan WS (2013) Time-shifted principal component analysis based cue extraction for stereo audio signals. In: Proceedings of ICASSP, Vancouver, Canada, pp 266–270
[HTG14] He J, Tan EL, Gan WS (2014a) Linear estimation based primary-ambient extraction for stereo audio signals. IEEE/ACM Trans Audio Speech Lang Process 22(2):505–517
[HGT14] He J, Gan WS, Tan EL (2014b) A study on the frequency-domain primary-ambient extraction for stereo audio signals. In: Proceedings of ICASSP, Florence, Italy, pp 2892–2896

[HGT15a] He J, Gan WS, Tan EL (2015a) Primary-ambient extraction using ambient phase estimation with a sparsity constraint. IEEE Sig Process Lett 22(8):1127–1131
[HGT15b] He J, Gan WS, Tan EL (2015b) Primary-ambient extraction using ambient spectrum estimation for immersive spatial audio reproduction. IEEE/ACM Trans Audio Speech Lang Process 23(9):1430–1443
[HGT15c] He J, Gan WS, Tan EL (2015c) Time-shifting based primary-ambient extraction for spatial audio reproduction. IEEE/ACM Trans Audio Speech Lang Process 23(10):1576–1588
[HPB05] Herre J, Purnhagen H, Breebaart J, Faller C, Disch S, Kjörling K, Schuijers E, Hilpert J, Myburg F (2005) The reference model architecture for MPEG spatial audio coding. In: Proceedings of 118th AES convention, Barcelona, Spain
[HKB08] Herre J, Kjörling K, Breebaart J, Faller C, Disch S, Purnhagen H, Koppens J, Hilpert J, Rödén J, Oomen W, Linzmeier K, Chong KS (2008) MPEG Surround-the ISO/MPEG standard for efficient and compatible multichannel audio coding. J Audio Eng Soc 56(11):932–955
[HPK12] Herre H, Purnhagen J, Koppens O, Hellmuth J, Engdegård J, Hilpert L, Villemoes L, Terentiv L, Falch C, Hölzer A, Valero ML, Resch B, Mundt H, Oh H (2012) MPEG spatial audio object coding—the iso/mpeg standard for efficient coding of interactive audio scenes. J Audio Eng Soc 60(9):655–673
[HHK14] Herre J, Hilpert J, Kuntz A, Plogsties J (2014) MPEG-H audio—the new standard for universal spatial/3D audio coding. J Audio Eng Soc 62(12):821–830
[HHK15] Herre J, Hilpert, J, Kuntz A, Plogsties J (2015) MPEG-H 3D audio—the new standard for coding of immersive spatial audio. J Sel Topics Sig Process 9:770–779
[HiD09] Hilpert J, Disch S (2009) The MPEG surround audio coding standard. IEEE Sig Process Mag 26(1):148–152
[Hol08] Holman T (2008) Surround sound up and running, 2nd edn. Focal Press, MA
[HWZ14] Hu R, Wang X, Zhao M, Li D, Wang S, Gao L, Yang C, Yang Y (2014) Review on three-dimension audio technology. J Data Acquis Process 29(5):661–676
[HKO04] Hyvärinen A, Karhunen J, Oja E (2004) Independent component analysis. Wiley, Hoboken
[Ios15] IOSONO (2015) http://www.iosono-sound.com/home/
[IrA02] Irwan R, Aarts RM (2002) Two-to-five channel sound processing. J Audio Eng Soc 50(11):914–926
[JHS10] Jeon SW, Hyun D, Seo J, Park YC, Youn DH (2010a) Enhancement of principal to ambient energy ratio for PCA-based parametric audio coding. In: Proceedings of ICASSP, Dallas, pp 385–388
[JPL10] Jeon SW, Park YC, Lee S, Youn D (2010b) Robust representation of spatial sound in stereo-to-multichannel upmix. In: Proceedings of 128th AES convention, London, UK
[JMG07] Jot JM, Merimaa J, Goodwin M, Krishnaswamy A, Laroche J (2007) Spatial audio scene coding in a universal two-channel 3-D stereo format. In: Proceedings of 123rd AES convention, New York, NY
[Ken95a] Kendall G (1995) A 3-D sound primer: directional hearing and stereo reproduction. Comput Music J 19(4):23–46 (Winter)
[KTT15] Kowalczyk K, Thiergart O, Taseska M, Del Galdo G, Pulkki V, Habets EAP (2015) Parametric spatial sound processing. IEEE Sig Process Mag 32(2):31–42
[LBP14] Lee T, Baek Y, Park YC, Youn DH (2014) Stereo upmix-based binaural auralization for mobile devices. IEEE Trans Consum Electron 60(3):411–419
[LBH09] Lossius T, Baltazar P, de la Hogue T (2009) DBAP–distance based amplitude panning. In: Proceedings of international computer music conference, Montreal, Canada
[MWC99] Martin G, Woszczyk W, Corey J, Quesnel R (1999) Controlling phantom image focus in a multichannel reproduction system. In: Proceedings of 107th AES convention, New York, NY

[MMS11] Melchior F, Michaelis U, Steffens R (2011) Spatial mastering-a new concept for spatial sound design in object-based audio scenes. In: Proceedings of international computer music conference. University of Huddersfield, UK
[MeF10] Menzer F, Faller C (2010) Stereo-to-binaural conversion using interaural coherence matching. In: Proceedings of 128th AES convention, London, UK
[MGJ07] Merimaa J, Goodwin M, Jot JM (2007) Correlation-based ambience extraction from stereo recordings. In: Proceedings of 123rd AES convention, New York
[Mil72] Mills W (1972) Auditory localization. In: Tobias JV (ed) Foundations of modern auditory theory. Academic Press, New York
[Moo98] Moore BCJ (1998) Cochlear hearing loss. Whurr Publishers Ltd., London
[PBD10] Plumbley M, Blumensath T, Daudet L, Gribonval R, Davies ME (2010) Sparse representation in audio and music: from coding to source separation. Proc IEEE 98(6):995–1016
[Pol05] Poletti M (2005) Three-dimensional surround sound systems based on spherical harmonics. J Audio Eng Soc 53(11):1004–1025
[Pot06] Potard G (2006) 3D-audio object oriented coding. PhD thesis, School of Electrical, Computer and Telecommunications Engineering, University of Wollongong. Available: http://ro.uow.edu.au/theses/539
[Pul97] Pulkki V (1997) Virtual sound source positioning using vector base amplitude panning. J Audio Eng Soc 45(6):456–466
[Pul07] Pulkki V (2007) Spatial sound reproduction with directional audio coding. J Audio Eng Soc 55(6):503–516
[PuK08] Pulkki V, Karjalainen M (2008) Multichannel audio rendering using amplitude panning [DSP Applications]. IEEE Sig Process Mag 25(3):118–122
[PKV99] Pulkki V, Karjalainen M, Välimäki V (1999) Localization, coloration, and enhancement of amplitude-panned virtual sources. In: Proceedings of 16th AES International conference spatial sound reproduction, Rovaniemi, Finland
[Ray07] Rayleigh L (1907) On our perception of sound direction. Philos Mag 13:214–323 (6th Series)
[ITU03] Rec. ITU-R BS.1284-1 (2003) RECOMMENDATION ITU-R BS.1284-1—General methods for the subjective assessment of sound quality
[ITU12] Recommendation ITU-R BS.775-3 (2012) Multichannel stereophonic sound system with and without accompanying picture, Geneva
[Rum02] Rumsey F (2002) Spatial quality evaluation for reproduced sound: terminology, meaning, and a scene-based paradigm. J Audio Eng Soc 50(9):651–666
[RZK05] Rumsey F, Zielicski S, Kassier R, Bech S (2005) On the relative importance of spatial and timbral fidelities in judgements of degraded multichannel audio quality. J Acoust Soc Am 118(2):968–976
[SAM06] Sawada H, Araki S, Mukai R, Makino S (2006) Blind extraction of dominant target sources using ICA and time–frequency masking. IEEE Trans Audio Speech Lang Process 14(6):2165–2173
[SBP04] Schuijers E, Breebaart J, Purnhagen H, Engdegard J (2004) Low complexity parametric stereo coding. In: Proceedings of 116th AES convention, Berlin, Germany
[SRK12] Schwarz A, Reindl K, Kellermann W (2012) A two-channel reverberation suppression scheme based on blind signal separation and Wiener filtering. In: Proceedings of ICASSP, pp 113–116
[SoE15] Sonic Emotion (2015) http://www2.sonicemotion.com/professional/
[SWR13] Spors S, Wierstorf H, Raake A, Melchior F, Frank M, Zotter F (2013) Spatial sound with loudspeakers and its perception: a review of the current state. In: Proceedings of IEEE 101(9):1920–1938
[StM15] Stefanakis N, Mouchtaris A (2015) Foreground suppression for capturing and reproduction of crowded acoustic environments. In: Proceedings of ICASSP, Brisbane, Australia, pp. 51–55

[SHT15] Sunder K, He J, Tan EL, Gan WS (2015) Natural sound rendering for headphones: integration of signal processing techniques. IEEE Sig Process Mag 32(2):100–113

[TaG12] Tan EL, Gan WS (2012) Reproduction of immersive sound using directional and conventional loudspeakers. J Acoust Soc Am 131(4):3215–3215

[TGC12] Tan EL, Gan WS, Chen CH (2012) Spatial sound reproduction using conventional and parametric loudspeakers. In: Proceedings of APSIPA ASC, Hollywood, CA

[ThP77] Theile G, Plenge G (1977) Localization of lateral phantom sources. J Aud Eng Soc 25(4):196–200

[TSW12] Thompson J, Smith B, Warner A, Jot JM (2012) Direct-diffuse decomposition of multichannel signals using a system of pair-wise correlations. In: Proceedings of 133rd AES convention, San Francisco

[UhH15] Uhle C, Habets EAP (2015) Direct-ambient decomposition using parametric wiener filtering with spatial cue control. In: Proceedings of ICASSP, Brisbane, Australia, pp 36–40

[UhP08] Uhle C, Paul C (2008) A supervised learning approach to ambience extraction. In: Proceedings of DAFx, Espoo, Finland

[UWH07] Uhle C, Walther A, Hellmuth O, Herre J (2007) Ambience separation from mono recordings using non-negative matrix factorization. In: Proceedings of 30th AES international conference, Saariselka, Finland

[UsB07] Usher J, Benesty J (2007) Enhancement of spatial sound quality: a new reverberation-extraction audio upmixer. IEEE Trans Audio Speech Lang Process 15(7):2141–2150

[VBG14] Vincent E, Bertin N, Gribonval R, Bimbot F (2014) From blind to guided audio source separation. IEEE Sig Process Mag 31(3):107–115

[Vir06] Virtanen T (2006) Sound source separation in monaural music signals," PhD Thesis, Tampere University of Technology, 2006

[VPS12] Välimäki V, Parker JD, Savioja L, Smith JO, Abel JS (2012) Fifty years of artificial reverberation. IEEE Trans Audio Speech Lang Process 20(5):1421–1448

[WaF11] Walther A, Faller C (2011) Direct-ambient decomposition and upmix of surround signals. In: Proceedings of IWASPAA, New Paltz, NY, pp 277–280

[WaB06] Wang DL, Brown GJ (eds) (2006) Computational auditory scene analysis: principles, algorithms, and applications. Wiley/IEEE Press, Hoboken

[WHO06] World Health Organization (2006) Primary ear and hearing care training resource. Switzerland

[Xie13] Xie BS (2003) Head-related transfer function and virtual auditory display, 2nd edn. Ross Publishing, USA

[YiR04] Yilmaz O, Rickard S (2004) Blind separation of speech mixtures via time-frequency masking. IEEE Trans Sig Process 52(7):1830–1847

[Zah02a] Zahorik P (2002) Assessing auditory distance perception using virtual acoustics. J Acoust Soc Am 111(4):1832–1846

[Zah02b] Zahorik P (2002) Direct-to-reverberant energy ratio sensitivity. J Acoust Soc Am 112 (5):2110–2117

[Zar02b] Zahorik P (2002) Direct-to-reverberant energy ratio sensitivity. J Acoust Soc Am 112 (5):2110–2117

[Zwi61] Zwicker E (1961) Subdivision of the audible frequency range into critical bands. J Acoust Soc Am 33(2):248

http://www.springer.com/978-981-10-1550-2

Spatial Audio Reproduction with Primary Ambient Extraction

He, J.

2017, IX, 132 p. 57 illus., 48 illus. in color., Softcover
ISBN: 978-981-10-1550-2