Acta Acustica 2022, **6**, 47
© The Author(s), Published by EDP Sciences, 2022
https://doi.org/10.1051/aacus/2022040

Available online at:
https://acta-acustica.edpsciences.org

Review Article Open Access

Spatial audio signal processing for binaural reproduction of recorded acoustic scenes – review and challenges

Boaz Rafaely[1,*], Vladimir Tourbabin[2], Emanuel Habets[3], Zamir Ben-Hur[2], Hyunkook Lee[4], Hannes Gamper[5], Lior Arbel[1], Lachlan Birnie[6], Thushara Abhayapala[6], and Prasanga Samarasinghe[6]

[1] School of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel
[2] Reality Labs Research, Meta, Redmond, WA 98052, USA
[3] International Audio Laboratories Erlangen (a joint institution of the Friedrich Alexander University Erlangen-Nürnberg (FAU) and Fraunhofer IIS), 91058 Erlangen, Germany
[4] Applied Psychoacoustics Laboratory (APL), University of Huddersfield, Huddersfield HD1 3DH, United Kingdom
[5] Audio and Acoustics Research Group, Microsoft Research, Redmond, WA 98052, USA
[6] Audio and Acoustic Signal Processing Group, The Australian National University, Canberra, Australian Capital Territory 2601, Australia

Received 25 March 2022, Accepted 8 September 2022

Abstract – Spatial audio has been studied for several decades, but has seen much renewed interest recently due to advances in both software and hardware for capture and playback, and the emergence of applications such as virtual reality and augmented reality. This renewed interest has led to the investment of increasing efforts in developing signal processing algorithms for spatial audio, both for capture and for playback. In particular, due to the popularity of headphones and earphones, many spatial audio signal processing methods have dealt with binaural reproduction based on headphone listening. Among these new developments, processing spatial audio signals recorded in real environments using microphone arrays plays an important role. Following this emerging activity, this paper aims to provide a scientific review of recent developments and an outlook for future challenges. This review also proposes a generalized framework for describing spatial audio signal processing for the binaural reproduction of recorded sound. This framework helps to understand the collective progress of the research community, and to identify gaps for future research. It is composed of five main blocks, namely: the acoustic scene, recording, processing, reproduction, and perception and evaluation. First, each block is briefly presented, and then, a comprehensive review of the processing block is provided. This includes topics from simple binaural recording to Ambisonics and perceptually motivated approaches, which focus on careful array configuration and design. Beamforming and parametric-based processing afford more flexible designs and shift the focus to processing and modeling of the sound field. Then, emerging machine- and deep-learning approaches, which take a further step towards flexibility in design, are described. Finally, specific methods for signal transformations such as rotation, translation and enhancement, enabling additional flexibility in reproduction and improvement in the quality of the binaural signal, are presented. The review concludes by highlighting directions for future research.

Keywords: Audio signal processing, Spatial audio, Virtual reality, Augmented reality, Array processing

1 Introduction

Binaural reproduction of acoustic scenes refers to the playback of sound at the listener's ears in a way that recreates a real-world listening experience of the scene. Ideally, the sound scene reproduced at another time and/or place should be perceptually indistinguishable from the real scene. Some important examples include capture and subsequent reproduction of musical performances or social events, as well as real-time video conferencing with immersive spatial audio.

Headphone-based playback of binaural sound, dating back to the 19th century [1], has become highly popular in recent decades with the availability of personal headphones. This also led to the rise in popularity of headphone-based binaural reproduction, and particularly, the reproduction of recorded acoustic scenes. The latter was initially based on binaural recording, using microphones placed at the ears of a manikin [2]. While providing an impressive spatial audio experience, binaural recording generally does not support listener individualization and

*Corresponding author: br@bgu.ac.il

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

head tracking, which are important for creating a realistic acoustic scene through headphone listening [3, 4]. The flexibility required for individualization and head-tracking was later obtained with the soundfield microphone and the Ambisonics spatial audio format [5]; these greatly advanced the recording and reproduction of real sound scenes through the separation of the recorded sound as captured by the microphone and the effect of the head on the signal at the ears, represented by the head-related transfer function (HRTF). Ambisonics was then extended to high-order Ambisonics [6–9] recorded by spherical microphone arrays [10, 11], providing higher spatial detail by supporting more recording channels. The seamless incorporation of HRTF into Ambisonics generated a remarkable listening experience within an elegant mathematical setting. Indeed, Ambisonics and HRTF have been the topic of extensive research in the past two decades, supporting a wide range of applications and research areas. For example, listening to sounds generated in simulated or measured acoustic spaces has been studied under auralization [12], investigating the listening experience from a human hearing perspective [13, 14]. The theory and practice of spatial audio recording and reproduction [15], and particularly Ambisonics [16], have been established, supported by advancements in spherical microphone array design and processing [17, 18]. New approaches to spatial audio processing and coding are still being proposed [19–21], facilitated by improved ways for headphone listening [22, 23]. However, in spite of these impressive advances over the past few decades, new emerging applications raise entirely new challenges for spatial audio in general, and binaural reproduction of recorded scenes, in particular.

A set of such emerging technologies that provides a new exciting platform for binaural reproduction applications is virtual reality (VR), augmented reality (AR), and mixed reality (MR) [12, 24, 25]. These originated from gaming, and have now been expanded to multimedia, education, personal communication, and virtual meetings, among many other areas. The new platforms introduce a unique set of challenges imposed by the fact that, in many cases, audio is captured by microphones that are embedded in consumer devices, which are often wearable. This is particularly challenging for the reproduction of recorded acoustic scenes. The first challenge is space and hardware limitations, which has led to the deployment of a small number of microphones of arbitrary arrangement, and often with unfavorable spatial diversity. Examples of these devices are mobile phones, laptops, smart speakers and VR headsets. These devices also introduce other challenges, imposed both by the motion of wearable and mobile arrays during signal acquisition [26], which hinders a stable listening experience of the reproduced scene, and by a low-latency constraint that occurs in applications involving real-time interactions, such as virtual video conferencing. In addition, acoustic scenes recorded by these devices may contain environmental noise and interfering sound, superimposed on the desired sound such as speech and music, which may degrade a virtual meeting, for example.

In synergy with the emerging technologies and applications, new directions in spatial audio signal processing are evolving that attempt to overcome the challenges mentioned above, and more. The aim of this review paper is to provide an updated account of these emerging methods, published in the past few years, and propose directions for future research. The paper first introduces a generalized framework for binaural reproduction of recorded acoustic scenes, then focuses on processing approaches, and concludes with prospects for future research. Regarding processing approaches, this paper first presents approaches that consider the microphone array as the dominant design element, and therefore require very specific microphone array designs. In binaural recording, two microphones are placed at the ears of a dummy head, while in Ambisonics, a dedicated array must be designed to capture spherical harmonics signals. In perceptual-based arrays, the microphones and their arrangement are by-design carefully configured to produce perceptually useful signals. Next, beamforming-based processing makes a step forward by lifting the constraints on array configuration, thus allowing a flexible design. Spatial filters, or beamformers, designed specifically for the array at hand, form the basis of the approach. This is then followed by parametric approaches, where a further step is made from array-focused methods to methods that exploit information in the sound field. The information is modeled and the model parameters are estimated, providing the basis for the spatial reproduction. Finally, machine- and deep-learning approaches provide an even more flexible framework that can exploit information both in the array configuration and in the sound field. Transformations such as rotation, translation and signal enhancement, tailored to the signal processing approaches, are then presented, followed by conclusions and an outlook for the future.

2 Overview

This section presents an overview of the entire process comprising binaural reproduction of recorded acoustic scenes. A generalized framework that encapsulates this process is first presented, from the acoustic scene being recorded to the perception and evaluation of the reproduced spatial audio. Each part of this process is reviewed in the following subsections, while processing approaches are reviewed in greater detail in the subsequent sections.

2.1 Generalized framework

The generalized framework of spatial audio signal processing for the binaural reproduction of recorded acoustic scenes is presented in Figure 1. The process presented in the figure starts from the *acoustic scene* – the real-world environment within which the sound is generated. This could be a concert hall with music sounds, an office with speech sounds, an outdoor environment with street sounds, and other scenes. A *recording* device, such as a microphone array of any type that is positioned in the scene, produces

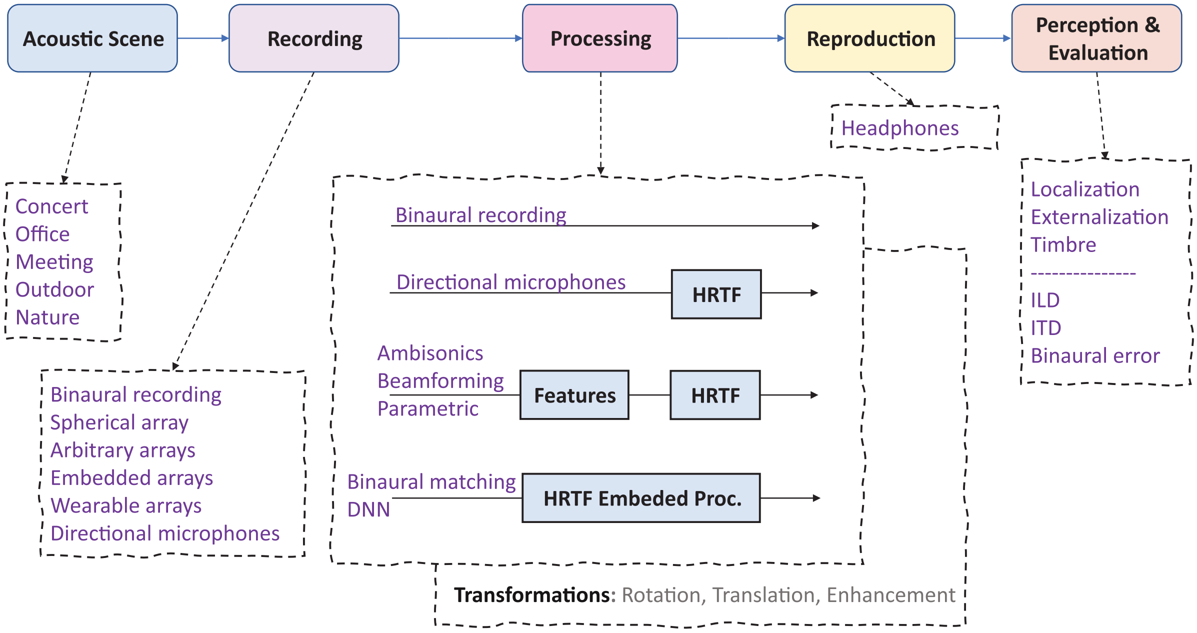

Figure 1. Generalized framework of spatial audio signal processing for binaural reproduction of recorded scenes.

recorded audio signals. The recording device can be anything from a dummy head directly recording binaural signals, to spherical arrays or arrays of other configurations. *Processing* is then applied to the recorded audio signals in preparation for reproduction; this stage is the main focus of this review paper and includes a wide range of spatial audio signal processing methods, from Ambisonics, through parametric audio, to deep learning. Note that Figure 1 shows another optional layer behind *processing* labeled *transformations*, which includes enhancement, rotation and translation. After processing, the spatial audio signal is ready for *reproduction* – this paper focuses on headphone reproduction, which is widely used in many applications. Finally, the headphone signals are perceived by listeners, or can be evaluated objectively; this is the final block of the framework, and is labeled *perception and evaluation*. More details on each block of the framework are presented in the following subsections.

2.2 Acoustic scenes

Spatial sound recording and binaural reproduction have found numerous applications in a large variety of acoustic scenes, ranging from relatively small indoor spaces to expansive outdoor areas. The indoor examples include offices and meeting rooms, where binaural reproduction has been employed for teleconferencing applications [27, 28]; these have recently received increased attention due to growing popularity of VR and AR platforms and the mushrooming of distance working/learning in response to the Covid pandemic. The acoustic source type of particular interest in this application is human speech. Another category of indoor acoustic scenes that has received significant attention in the past few decades is concert halls. The applications include recordings of music or other artistic performances [29], and perceptual assessment and comparison of concert hall sound [30, 31]. These applications are usually characterized by elevated reverberation, and the acoustic sources of interest are primarily musical instruments and human voices. Multiple outdoor applications have also been explored. For example, spatial recordings have been utilized to capture urban sounds, including traffic, subway stations, and social gatherings; these were used to facilitate perceptual soundscape studies using various reproduction methods [32, 33]. Finally, spatial sound recording methods have also been proposed for use in open outdoor environments to record nature sounds like waterfalls, birds, and wind [34]; these methods were utilized in applications related to art and entertainment [35].

2.3 Recording devices

A large variety of devices have been successfully employed for spatial sound capture. The function of the capture devices is to record the essential spatial information that enables either physically accurate [33] or perceptually plausible [36] reproduction of the signals at the listener's ears. Probably, the most straightforward recording device enabling binaural reproduction is the binaural microphone (see, for example, [37]), which can be placed on the head of a human subject [32] or on an acoustically designed binaural fixture [2, 38]. More complex microphone array systems have been proposed to improve spatial capture resolution and facilitate sound field manipulation. These include the B-format soundfield microphone array (comprised of four capsules located on the faces of a tetrahedron [39]), high-order spherical arrays [40] (that facilitate sound field decomposition and manipulation in the spherical harmonics domain [41, 42]), approaches that support flexible recording arrays [43, 44], and very large microphone-array systems with interpolation processing [45]. There also exist various perceptually-motivated microphone arrays (PMMAs), designed for capturing acoustic scenes. Whilst high-order arrays attempt to reconstruct the sound field in the reproduction process in a way that is as physically accurately as possible, perceptually motivated arrays focus on plausibly representing the sound field using psychoacoustic cues such as interchannel time- and level-differences and interchannel coherence [36]. The capture devices mentioned above enable various processing methods for enhancing and manipulating the sound field prior to reproduction, as described in Section 3.

2.4 Processing

The processing block in Figure 1 transforms the recorded signals from the previous block into binaural signals ready for headphone reproduction in the following block. This aim can be achieved with a wide range of approaches and methods, from binaural recording, which directly produces a binaural signal, to methods such as Ambisonics and beamforming-based processing, which employ microphone arrays and more complex operations. This variety of methods is reviewed in more detail in the following sections, constituting the main part of this review paper. The methods include binaural recording, Ambisonics, perceptually motivated approaches, parametric processing, beamforming-based processing, machine- and deep-learning based methods, and transformations such as signal enhancement, translation and rotation of the listener's virtual position.

2.5 Reproduction

The reproduction block in Figure 1 converts the binaural signals back to sounds using electroacoustic transducers. When dedicated transducers are used for the left and right ears, there is no cross-talk between the left binaural signal and the right ear and vice versa, which allows for more direct control of the sound at the ears. The most common device for binaural playback of sound is the headphone [4, 46], which comes in different forms, including circum-aural (over-the-ear), supra-aural (on-the-ear), earbud, in-ear, and bone-conducting. Some over-the-ear headphones use an open design, allowing audio leakage out of the earpieces and ambient sound leakage into the earpieces. Other headphones use a closed design to preclude leakage. When using headphones to play binaural signals, even though the real sound sources are the electroacoustic transducers at the ears, sounds can still be perceived outside the listener's head by carefully controlling the left and right ear signals. This phenomenon, known as sound externalization, contributes to the realistic perception of a virtual scene [47].

Another factor related to headphone reproduction that contributes to realistic perception is head-tracking, which stabilizes the perceived virtual scene, despite the listener's head movements [3, 48, 49]. Head-tracking requires dedicated hardware, such as a head-mounted inertial measurement unit that operates in real time with limited latency [50, 51]. Finally, the frequency response of the headphone may affect perception, and so often this response is compensated for using headphone equalization [52–54].

2.6 Perception and evaluation

The last block of the generalized framework presented in Figure 1 is perception and evaluation. Perception is the aim of the binaural reproduction process – to recreate spatial sounds that are perceptually indistinguishable from the real sounds, i.e., the listener perceives the spatial sound authentically as if he/she were actually in the scene [55]. Therefore, evaluating whether this aim has been achieved is of fundamental importance. The evaluation can be both technical, by means of errors in the reproduced binaural signals, and perceptual, by listening tests. Technical evaluation can be performed, for example, by quantifying the errors between a reference signal and the reproduced signal, or by evaluating the accuracy of binaural cues, such as interaural time- and level-differences (ITD and ILD), and interaural cross-correlation (IACC) [56–60].

Perceptual evaluation has traditionally used a global attribute called "basic audio quality" or subjective preference. However, recent studies in spatial audio increasingly tend to evaluate different systems in terms of specific attributes (e.g., [61–66]). Examples of such attributes are sound source *localization*, *externalization*, *coloration*, *apparent source width* (*ASW*) and *listener envelopment* (*LEV*). More general measures for evaluating the overall perceptual accuracy, such as *plausibility* [67, 68] and *authenticity* [55, 69], have also been suggested. Once listening tests have been performed, comprehensive analysis could lead to a perceptual model to replace further listening tests. Examples include localization and externalization models [70–74], and a surround sound quality model [75]. Moreover, machine-learning algorithms have also been suggested for the evaluation of spatial perception [76, 77]; this will be further discussed in Section 3.6.

While the auditory attributes stated above have traditionally been studied in the context of a static listener position and head orientation with a fixed perspective, recent developments in VR and AR require 6-degrees-of-freedom (6DoF), where the listener is free to rotate his/her head and also walk around in a virtual or real space. New tools have been developed to perform listening tests and behavioral studies in interactive virtual environments [78]. In a recent study, various direct and indirect audio quality evaluation methods were compared in virtual reality scenes of varying complexity [79]. It was found that *rank-order elimination* proved to be the fastest method, required the least amount of repetitive motion, and yielded the highest discrimination between spatial conditions. Scene complexity was found to be a main effect within results, while behavioral and task load index results imply more complex scenes, and interactive aspects of 6-DoF VR can impede quality judgments. Recent perceptual studies [80, 81] also found that such a dynamic environment could lead to dramatic changes in the perceived reverberation, loudness, ASW and LEV, making evaluation much more challenging under such dynamic conditions.

3 Processing approaches

This section presents a review of methods associated with the *processing* block in Figure 1, providing the mapping from the captured microphone signals to binaural signals ready for listening.

3.1 Binaural recording

In binaural recording, microphones are placed at the ears of a dummy head, capturing the sound at the ears of a potential listener at the recording position. While binaural recordings have a long history [1], they are still widely used today, as they generate binaural signals, ready for listening, without the need for further processing [4]. While an attractive option in spatial audio, binaural recording suffers from two main limitations, both related to the innate embedding of the HRTF in the recording. The first is that head-tracking is typically not possible, as the head position is captured in the recording. The second is that individualized HRTF cannot be supported, as the signal embeds the HRTF of the dummy head. Solutions to the former exist, such as motion-tracked binaural recordings [82, 83], or binaural cue adaptation [84]; however, these are still limited in their accuracy and flexibility. These two limitations call for more flexible recording solutions, in which the sound field is recorded separately from the HRTF, which can then be integrated in post-processing. Such approaches are presented next.

3.2 Ambisonics

Ambisonics was first introduced in the 1970s as a way to record and reproduce spatial audio using 4 audio channels, denoted as first-order Ambisonics (FOA) [5, 85–87]. Around the late 1990s, the higher-order Ambisonics (HOA) technology, using a spherical harmonics formulation, emerged [6, 8, 9]. FOA and HOA were originally developed for loudspeaker array reproduction. In 1999, an approach for headphone reproduction of Ambisonics signals was introduced [88], using "virtual loudspeaker reproduction". Headphone reproduction using Ambisonics has been significantly advanced in the past decade as new applications have emerged (see Sect. 1). Specifically, a formulation in the spherical harmonics domain of binaural reproduction using Ambisonics signals, which also employed a spherical harmonics representation of the HRTF [89], was presented [7, 42, 90, 91]. The use of the spherical harmonics formulation has become popular in recent years, due to the possibilities for efficient processing in the spherical harmonics domain, the inherent separation of the sound field and the HRTF representations, and the ease of rotation of these representations, which is useful for head-tracking, for example [16, 92–95]. Figure 2 presents a general diagram for Ambisonics-based binaural reproduction, showing how the spherical harmonics representations of the sound field and the HRTF are combined to form the binaural signal.

HOA signals can be derived from microphone recordings, typically using a spherical array such as the 4th order Eigenmike [96]. The process of computing the Ambisonics signals is often termed plane-wave decomposition (PWD) [97], because Ambisonics can be related to the plane-wave amplitude density function [18]. However, practical arrays have a limited number of microphones, which may limit the spherical harmonics order and the spatial resolution, and introduces spatial aliasing at high frequencies [98]. Methods that reduce aliasing may extend the frequency range of operation of the array, for example, by aliasing cancellation [99]. Moreover, the typically-small array size affects the robustness of PWD at low frequencies due to the low magnitude of the radial functions that encode scattering off the array [97]. A robust PWD method was recently proposed to overcome these low frequency limitations [100]. Another approach to enhance the Ambisonics signals is by upscaling, which aims to extend the spherical harmonics order, and leads to enhanced spatial resolution and higher-quality spatial audio signals. Earlier work includes the employment of compressed sensing [101–103] and sparse decomposition based on dictionary learning [104], while more recent work includes the employment of sparse recovery [105] and deep-learning [106–108]. Order-limited Ambisonics signals translate to order truncation of the HRTF [109], which may have a detrimental effect on the perception of the reproduced binaural signals [93, 110]. Several methods that overcome this limitation have been suggested in recent years [94]. Correction of spectral deficiencies by diffuse-field equalization was suggested in [110, 111]. Other approaches suggested modifying the HRTF phase component, e.g., time-aligned binaural decoding [95], magnitude least-square (MagLS) [112], and bilateral Ambisonics [56]. The phase was shown to contribute significantly to the increased order of the HRTF [113], and so its modification leads to improved reproduction using low-order Ambisonics.

Ambisonics has been established as a common standard for spatial audio, but even with the improvements described above, it has limitations that drive the search for improved solutions. A main limitation appears when Ambisonics with a low spherical harmonics order is used, for which the binaural reproduction may be of poor quality. Other limitations are detailed next. The frequency range of the Ambisonics signals when captured with compact microphone arrays such as a spherical array may be limited by spatial aliasing and robustness constraints, as discussed earlier in this section. On the positive side, Ambisonics readily supports spatial rotation, which is useful for head-tracking and 3 degrees-of-freedom (3DoF) rendering. However, the incorporation of spatial translation is not trivial [114]. Another limitation is that the recording of Ambisonics signals often requires a spherical array, which may not be available when using microphone arrays embedded in consumer devices, for example. Finally, the recording of real scenes may also be corrupted by noise and interference and may require enhancement. Various methods that try to

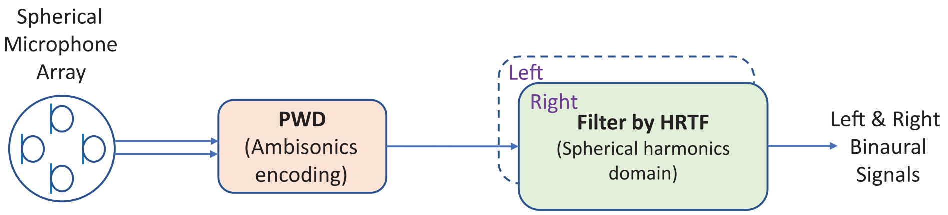
Spherical Microphone Array
PWD
(Ambisonics encoding)
Left
Right
Filter by HRTF
(Spherical harmonics domain)
Left & Right Binaural Signals

Figure 2. A block diagram illustrating Ambisonics-based spatial audio processing, showing a spherical microphone array and operations of plane wave decomposition (PWD) and filtering by HRTF.

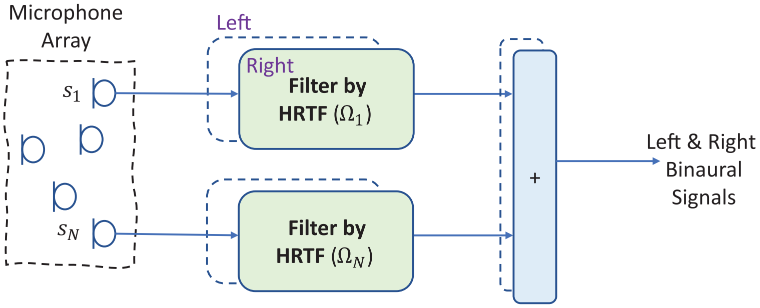
Microphone Array
s_1
s_N
Left
Right
Filter by HRTF (Ω_1)
Filter by HRTF (Ω_N)
+
Left & Right Binaural Signals

Figure 3. A block diagram illustrating perceptually motivated microphone-array processing for binaural reproduction.

overcome these limitations of Ambisonics are described in the next sections.

3.3 Perceptually motivated approaches

As outlined in Section 2.3, PMMAs aim to preserve psychoacoustic cues directly in the microphone-array signals, such that perceptual attributes of the acoustic scene are plausibly rendered. This is in contrast to reconstructing the sound field in a physically accurate manner in post-processing, an approach often employed when Ambisonics signals are computed from spherical microphone arrays, for example, as reviewed in Section 3.2. In particular, most PMMAs focus on manipulating interchannel time difference (ICTD), interchannel level difference (ICLD) and interchannel correlation (ICC) for virtual image localization and spatial-impression rendering. The concept relies on the perceptual phenomena of summing localization and the precedence effect [69]. Typically, the signals of a PMMA do not require any further decoding process for reproduction; each microphone-array signal is discretely routed to each corresponding loudspeaker. For binaural reproduction using a PMMA recording, loudspeakers are replaced by virtual sources, while the source signals are convolved with the head-related impulse responses (HRIRs) associated with the virtual source positions. This approach, illustrated in Figure 3, offers an attractive advantage – binaural reproduction with good perceptual quality can be achieved even with a small number of microphones.

There exist several models of the ICTD and ICLD trade-off for controlling the degrees of image shift [115–117], that are used for designing the spacing and relative angle between microphones in an array. These models can also be used to affect the characteristics of a virtual source for a given perceived source position. In particular, higher ratios of ICTD to ICLD lead to more spacious, but less localizable, sources, and a greater sense of depth and spread [63]. Achieving a sufficient amount of interchannel decorrelation is another important design goal for PMMAs. Decorrelation is not only important for an auditory spatial impression, i.e., ASW and LEV, [58, 118], but also for extending the size of the listening area in loudspeaker reproduction. This is of less importance in binaural reproduction, where the listener is always at the sweet spot [119, 120]. Decorrelation is also frequency dependent [121]. Since low-frequency decorrelation has been reported to be important for LEV [122], various decorrelation methods have been proposed [118, 123, 124]. Furthermore, decorrelation of vertically oriented signals has been found to have a minimal, or no, effect on the vertical spread of virtual sources, depending on source frequency [124, 125]. This allows a three-dimensional microphone array to be more compact vertically. Examples include the ORTF-3D [126] and ESMA-3D [59] arrays.

Despite providing good perceptual quality with a small number of microphones, PMMAs do not directly support generic representations like Ambisonics, making this approach specific to a loudspeaker configuration. With this limitation in mind, methods have been developed to transform PMMA signals into Ambisonics. A recent study [127] investigated the perceived spatial and timbral degradation when signals of various PMMAs were directly encoded to Ambisonics with different orders, and binaurally reproduced using the MagLS decoding method [95]. A multiple stimulus with hidden reference and anchor (MUSHRA) listening test revealed that the perceived degradation was minimal with the order of 2 or higher, depending on the decoder. This suggests that Ambisonics could be a useful coding and delivery format for PMMA recordings.

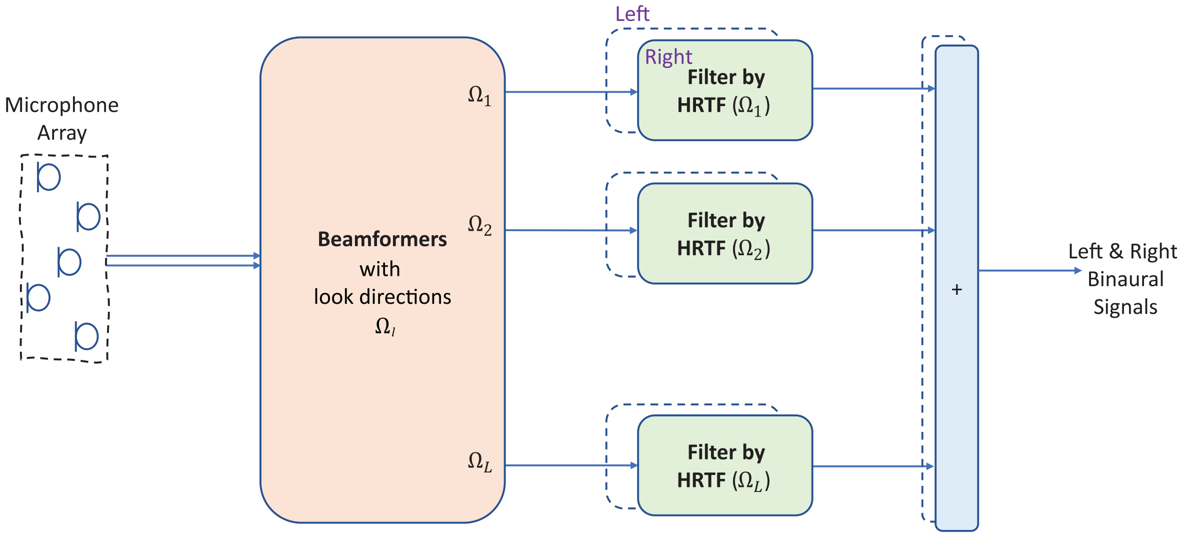
Microphone Array
Beamformers with look directions Ω_l
Ω_1
Ω_2
Ω_L
Left
Right
Filter by HRTF (Ω_1)
Filter by HRTF (Ω_2)
Filter by HRTF (Ω_L)
+
Left & Right Binaural Signals

Figure 4. A general diagram illustrating beamforming-based spatial audio processing.

In summary, PMMAs aim for high perceptual quality with a small number of microphones, but come at the cost of highly specific microphone-array designs. Alternative approaches with a similar aim, but supporting more flexible array designs are reviewed next.

3.4 Beamforming-based processing

Beamforming-based processing refers to the family of methods that transform microphone signals into a binaural signal in two stages. In the first stage, beamforming, or spatial filtering, is applied to the microphone-array signals, most commonly to represent sound field components associated with specific directions. Then, in the second stage, these components are filtered by the appropriate HRTF and combined to form binaural signals, as illustrated in Figure 4. This is a useful approach, and in its current form it has been developed with great flexibility to array configuration. Ambisonics signals derived from spherical microphone arrays can be considered as a special case of this approach, as detailed below.

Early work developed within this framework employed Ambisonics signals, and did not explicitly use the term beamforming. Here, Ambisonics signals were decoded into signals that are used to directly drive an array of actual loudspeakers [128], or, alternatively, an array of virtual loudspeakers [42]. The set of virtual loudspeaker array signals was further filtered with HRTF to produce binaural signals. This approach was later extended from Ambisonics to spherical arrays in general, by decomposing the measured microphone signals into spherical harmonics and then plane waves, finally reproducing binaural signals by combining each plane wave with the appropriate HRTF [41]. Further work implemented this approach on a real spherical array [129], and analyzed the approach theoretically [130].

Having established the generation of virtual loudspeaker signals and then signals related to PWD, the approach was then extended mathematically to employ beamformers to estimate signals in specific arrival directions. This approach builds on a well-established theory of beamforming [131], with well-defined design methods. Early work incorporated maximum-directivity beamformers, leading to Ambisonics signals and PWD for spherical arrays [132–134]. Later, other beamformers, such as the delay-and-sum beamformer, were also investigated [135]. However, these studies were limited to spherical arrays.

Another direction of research work related to beamforming that was also applied to spherical microphone arrays used beamforming or spatial filtering to shape the directivity of the sound field, thus reducing noise arriving from directions attenuated by the spatial filter (see Sect. 4.2), with the entire process embedded in an Ambisonics setting [136–139]. This approach demonstrated a trade-off between noise reduction and spatial audio quality. A different approach, also related to the methods in Section 4.2, placed emphasis on noise reduction using high-performance beamformers, such as the maximum-directivity distortionless response (MVDR) beamformer [140] and the linearly-constrained minimum variance (LCMV) beamformer [141]. These approaches only partly supported spatial audio reproduction quality by incorporating constraints in the beamformer design to ensure basic cues of the binaural signals, such as ILD and ITD for specific sources at the beamformer output. This approach did not involve HRTF and the quality of the reproduced spatial audio was limited.

In more recent studies, the beamforming approach developed in previous work was applied to arrays of arbitrary configuration, such as arrays mounted on helmets [142], or glasses [143], linear arrays [43, 144], and wall mounted planar arrays [145]. These recent studies extended previous work which was mostly developed for Ambisonics signals. Design methods were further developed by proposing a framework for selecting the number of beamforming directions [145], by direct matching of microphone signals to binaural signals [44, 143], and by designing virtual artificial heads [43, 144]. The last may require efficient representations of HRTF, e.g., [146]. These are initial steps in the development of methods that will support high quality binaural reproduction based on practical microphone arrays, such as wearable arrays and arrays with arbitrary configuration.

In summary, while considerable progress has been made for beamfroming-based binaural reproduction, most previous work was developed for Ambisonics signals; it may not be possible to accurately compute these signals from signals measured by arrays with a small number of microphones (e.g., from microphones mounted on devices). For such arrays, current beamforming-based design methodology may offer an attractive and flexible alternative; however, at this point in time, further research providing theoretical grounding is required, as well as further development of processing methods to support high quality binaural reproduction from such arrays.

3.5 Parametric processing

Parametric processing is based on relatively simple, in some cases, perceptually motivated, sound field modelling. The processing generally consists of two steps. In the first step a specific sound field model is assumed and its parameters and signals are estimated, while in the second step the binaural signals are synthesized. Reproduction based on a small number of parameters may be advantageous when the complexity of the sound field cannot be captured by the recording array. In this case, estimating a small number of perceptually important parameters may be more useful than attempting to capture the full complexity of the sound field.

One of the earliest approaches of parametric signal processing for spatial audio is based on decomposing the sound field into a direct-sound component, representing the sound source, and a diffuse sound component, representing reflections and room reverberation. The approach, referred to as DirAC (directional audio coding) [147], was developed for FOA. A similar approach decomposed the sound field into primary and ambient components [148]. The former component is highly correlated between input channels (representing sources), and the latter are uncorrelated (representing reverberation and background noise). Both approaches process the signals in the time-frequency domain, exploiting the sparsity property of audio signals such as speech. Therefore, while only one source per time-frequency bin is modeled, overall, these approaches can model an acoustic scene with multiple sources. Another alternative, high-angular-resolution plane-wave expansion (HARPEX) [149, 150], models two plane waves per time-frequency bin, complemented by two opposing plane waves, thus enriching the plane-wave model.

While useful, these early approaches for parametric spatial audio processing are limited due to their simplistic models [21], and so methods employing more complex models have been developed [20, 151, 152]. With the aim of extracting multiple dominant plane waves from complex sound fields, sparse recovery approaches have been employed [20, 153]. Multiple plane-wave modeling and a more flexible representation of the reverberant part of the sound field have also been the basis for HOA extensions of DirAC [20, 154–156], leading to improved spatial resolution and a more accurate representation of complex sound fields. This approach, developed for Ambisonics signals and spherical arrays, has been extended to incorporate general microphone arrays, by employing optimal multiple channel filters to estimate direct signals from sources [20, 152, 157]. Figure 5 presents a general block diagram, capturing the main processing blocks common to parametric spatial audio signal processing for binaural reproduction. While the approaches discussed above are often presented in the context of loudspeaker reproduction, they are nevertheless relevant for headphone reproduction by employing virtual loudspeakers, or by rendering sources by incorporating HRTF [20, 21].

Overall, parametric processing has been a promising avenue for binaural reproduction from microphone-array recordings, as it has the potential to capture important spatial information through the modelling process. Further research may provide high-quality reproduction even with challenging environments that include multiple dynamic sources, spatially complex sources [158, 159], reverberation and noise, and by employing compact arrays with only a few microphones. Improved methods for estimating information on individual sources and on reverberant components, as well as methods that incorporate early room reflections [160–163], may advance the parametric approach even further. The parametric processing approach also supports signal transformations such as rotation and translation, due to the simplified sound-field representation, as will be further discussed below.

3.6 Machine- and deep-learning based processing

With the advent of deep-learning methods, machine learning has seen broad application for a wide variety of research problems, including in the fields of audio and acoustics. Recently, novel machine-learning-based methods that fit within the generalized framework shown in Figure 1 have been proposed.

Understanding the characteristics of the acoustic environment may be useful in a spatial audio processing framework (see Sect. 2.2). The annual IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE) [164] includes contributions related to automatically classifying the type of acoustic scene [165], or detecting and localizing sound events from spatial audio recordings [166]. Grumiaux proposed the use of the time-domain velocity vector as an input feature for a deep neural network (DNN) to count and localize multiple speakers in Ambisonics signals [167]. A related problem is the blind estimation of room acoustic parameters from audio recordings [168], including the estimation of reverberation time and the early-to-late reverberation ratio [169–173].

Given a recording of an acoustic scene, data-driven and machine-learning approaches can be used for audio

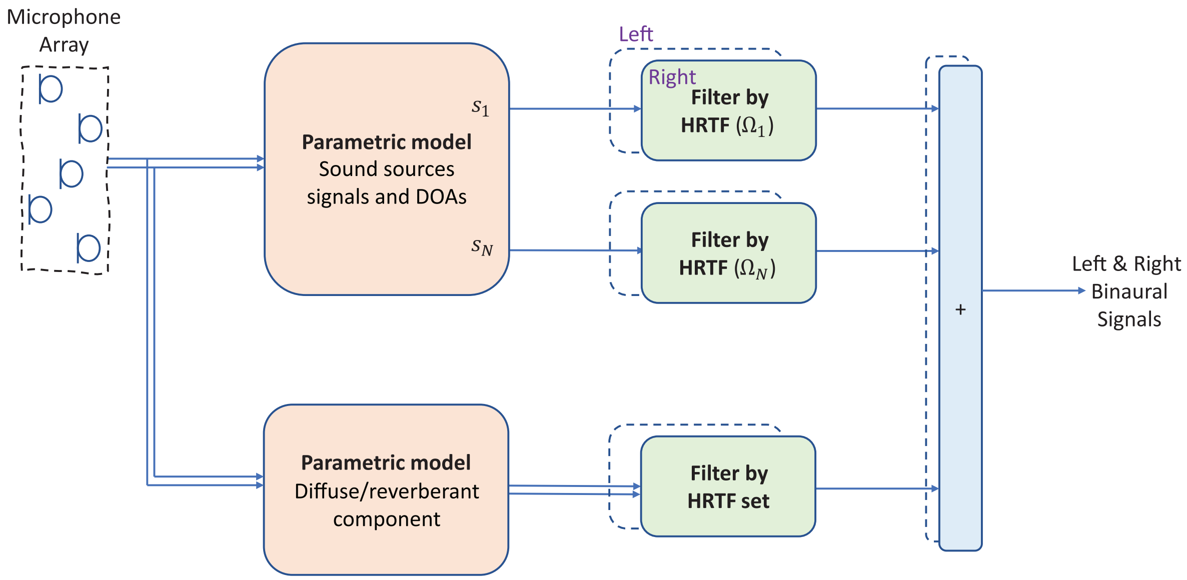
Microphone Array
Parametric model
Sound sources signals and DOAs
s_1
s_N
Parametric model
Diffuse/reverberant component
Left
Right
Filter by HRTF (Ω_1)
Filter by HRTF (Ω_N)
Filter by HRTF set
+
Left & Right Binaural Signals

Figure 5. A general diagram illustrating parametric-based spatial audio processing, incorporating a first stage of parametric modeling and parameter estimation, and a second stage of HRTF-based binaural reproduction.

processing (see Sect. 2.4). A method has been proposed in [174] for upmixing monophonic recordings to FOA by combining audio processing and computer-vision methods to infer sound source locations from a panoramic video recording of the scene. The estimated Ambisonics signals can then be processed further for binaural reproduction. Directly deriving the binaural output signals from a monophonic recording has also been proposed, taking into account the position and orientation of the listener relative to the source [175]. In another study, convolutional neural networks were employed to upscale the Ambisonics order of encoded FOA recordings [106]. On the reproduction side, general adversarial networks were proposed to reduce the error when rendering Ambisonics-encoded sound fields over four loudspeakers [108]. Finally, data-driven and machine-learning-based approaches have been proposed for the perceptual evaluation of reproduced scenes. A model that predicts front-back and elevation perception of sound sources was introduced in [70], while predicting spatial audio quality using computational models was proposed in [77]. For an extensive review of current data-based spatial audio methods the reader is referred to the work by Cobos et al. [176].

With the increased popularity of machine- and deep-learning research, it is expected that these approaches will play an increasingly more significant role in the near future – for spatial audio, in general, and for the binaural reproduction of recorded acoustic scenes, in particular. As these approaches are data-driven, they have the potential to overcome limitations imposed by microphone-array configurations, and to implicitly exploit information embedded in the sound field, leading to highly flexible solutions; nevertheless, these solutions may require tailoring to specific systems and applications.

4 Transformations

This section presents processing methods that can be considered as additions to the main processing chain of mapping microphone signals to binaural signals, as illustrated in Figure 1. These include signal enhancement to reduce unwanted interfering sounds in the spatial audio signal, and translation and rotation that support the mobility of a listener in a virtual audio environment.

4.1 Rotation and translation

During binaural reproduction with headphones, listeners may rotate their heads, leading to a corresponding rotation of the acoustic scene, which is perceived as unnatural. This can be corrected by head-tracking, i.e., rotating the acoustic scene to counter the listener's head rotations, thereby stabilizing the virtual scene and providing the feeling of immersion in a real scene. This head-tracking is denoted as having 3DoF. Furthermore, listeners may move freely, i.e., walk through the reproduced scene with a combination of rotational and translational movements. The latter refers to moving forwards and backwards, up and down, and left and right. The translation is often referred to as sound field translation, sound field navigation, or scene walk-through. When paired with rotation, the complete freedom of movement is denoted 6DoF. The objective of 6DoF reproduction is to enable a listener to walk through an acoustic scene in VR/AR, leaning close to sound sources or reflectors and hearing a realistic life-like recreation of the true experience (ideally with matched visuals).

A schematic illustrating how recordings are compensated for listener rotation and translation for the case of an Ambisonics signal is given in Figure 6. Typically, the

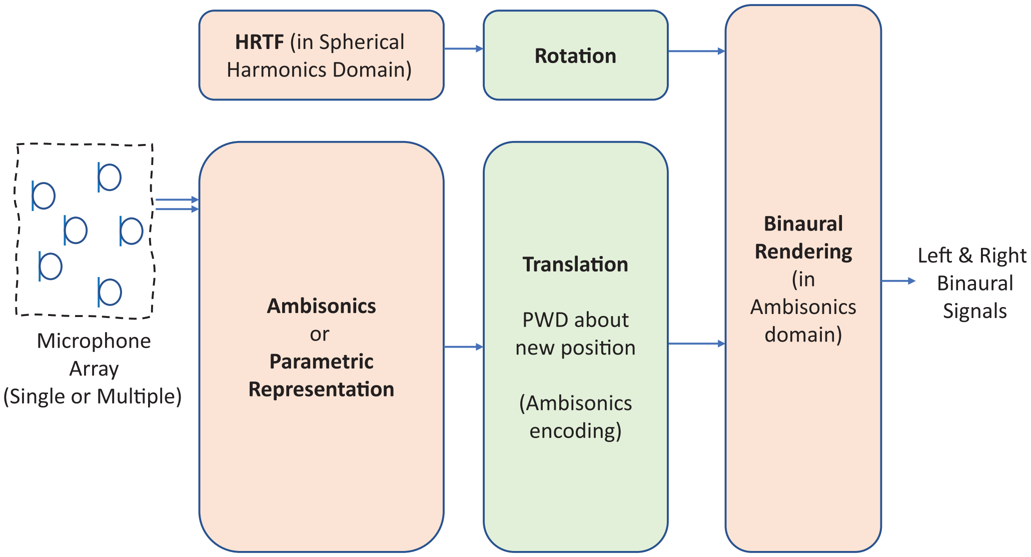
HRTF (in Spherical Harmonics Domain)
Rotation
Microphone Array (Single or Multiple)
Ambisonics or Parametric Representation
Translation
PWD about new position
(Ambisonics encoding)
Binaural Rendering (in Ambisonics domain)
Left & Right Binaural Signals

Figure 6. Generalized framework of sound field translation and rotation for binaural reproduction of recorded scenes.

recorded sound field is processed into an intermediate representation stage that supports sound field rotation and translation. The intermediate sound field is then recomposed back to an Ambisonics representation at the listener's new position, and the binaural signals are rendered as usual. Further detail on the approaches for listener rotation and translation are provided in the following.

A straightforward method for enabling head rotation is to record the scene with multiple binaural microphones at different azimuth rotations. For example, by having microphones [82, 177] or binaural microphones [178] placed around the equator of a sphere. During reproduction the listener's head rotation is tracked and the microphone signals closest to the ears are interpolated or directly played back. However, currently, head rotation for 3DoF is more commonly achieved by rotating the Ambisonics representation of the sound field [128, 179], or equivalently, rotating the Ambisonics representation of the HRTF [180]. Ambisonics rotation is easily performed by applying a time- and frequency-independent rotation matrix to the Ambisonics coefficients [181–186]. The challenges of head-rotation-enabled binaural reproduction for recording by non-standard or wearable microphone arrays [187–189] are still the subject of ongoing research.

There are three main approaches towards enabling 6DoF, each distinguished by the recording setup. The first is a source-based approach, where spot microphones are used to record each sound source individually within the scene [190]. The recorded scene is virtually pieced back together by representing the sources as virtual objects at similar positions. The virtual object signals are panned and amplified depending on the listener's real-time position and rotation. This approach is easy to adapt to different binaural rendering methods. However, no source-directivity information is captured, and the specific acoustics of the environment are not typically captured or reproduced.

The second 6DoF technique, denoted extrapolation-based, records the acoustic scene from a single spatial position, usually with a single HOA microphone. The Ambisonics recording is processed into a secondary representation built out of virtual loudspeakers [191–193], virtual microphones [194], virtual near-field point-sources [195–198], virtual far-field plane waves [199–202], or virtual near- and far-field sources [203]. Alternatively, the HOA recording can be directly re-expanded about a translated position without the secondary representation [182, 204]. The extrapolation-based translation, however, is usually limited by a sweet-spot distance that is defined by the well-known truncation properties of the Ambisonics decomposition [205]. To address this limitation, methods often use additional assumptions or parametric information about the recording to gain extended translation. For example, methods use known or estimated source directions and/or distances [192, 196, 198, 200, 206], a distance map [197, 199], or spatial sparsity assumptions [203, 207]. While the translation distance is limited, the extrapolation-based approach benefits from its unobtrusive and cost efficient use of a single HOA microphone. Lastly, the potential applicability to other single recording devices, such as wearable microphone arrays, suggests that the extrapolation approach will continue to develop.

The third 6DoF technique, denoted interpolation-based, records the scene from multiple spatial positions with a distributed grid of Ambisonics (first-order or higher-order) microphones. Existing approaches for Ambisonics interpolation can broadly be classified into two categories: parametric approaches in the time-frequency domain, and broadband approaches in the time domain. The parametric approach exploits time-frequency analysis of the multiple Ambisonics recordings to infer underlying source characteristics (mainly the location information), which are then explicitly [208–212] or implicitly [213–215] used to render the reproduced sound field at interpolated listening positions. Tracking-based solutions for moving sources have also been proposed [216–220]. Additional information on source locations enlarges the supported range of shifted listening perspectives with high spatial definition, yet the time-frequency processing often results in musical noise artifacts. In contrast, broadband approaches such as weighted averaging and virtual loudspeaker objects (VLO) make no attempt at analyzing underlying source characteristics, and their time-domain processing avoids the risk of introducing musical noise. The weighted averaging method [221–223] applies distance-based weights to each recording and has a few notable shortcomings, including a limited listener movement region and poor localization accuracy. In contrast, the VLO method [193, 224] maps the recordings to multiple surround playback rings of virtual loudspeaker objects, whose direction and amplitude vary with the desired listener position, thus providing enhanced spatial fidelity. More recently, in [225, 226], the authors presented methods that merge and extend the concepts of parametric and broadband interpolation. Overall, interpolation-based approaches offer potentially longer translation distances, but with the trade-off of the increased costs associated with using multiple HOA microphones [227, 228].

While numerous rotation and translation solutions have been recently developed, capturing and accurately reproducing large acoustic scenes still remains an open problem. Future directions include the extension of these methods to more general signals beyond Ambisonics, and improving the accuracy and translation regions to support realistic free walking in virtual and augmented reproductions of captured sound scenes.

4.2 Signal enhancement

Spatial audio signals may be composed of both desired components, such as speech and music, and undesired components, such as noise and interfering sounds. Therefore, in addition to processing aiming at binaural reproduction, signal enhancement may also be required in order to attenuate the interfering components, thereby delivering to the listener high quality spatial audio which is also clean.

This problem has been investigated for hearing aids, where the delivery of clean speech is of great importance, while binaural hearing aids also aim to deliver spatial cues to the listener. With this in mind, binaural beamformers that aim to attenuate undesired signal components [140, 141] have been developed; these were also extended to include time-frequency masking [229]. However, because spatial information relies on beamforming constraints, it is only partially preserved in the binaural signal. Furthermore, these methods are designed for binaural microphone arrays and may not always be applicable to general arrays.

With the aim of overcoming the limitations of binaural signal enhancement, several studies developed enhancement solutions for Ambisonics signals. In the first approach, directional constraints were introduced into the Ambisonics encoding process to attenuate directional interferences. Then, with the aim of affording more flexibility to target noise fields that are not highly directional single sources, a directional shaping filter that allocates higher directional gain to directions with higher signal-to-noise ratio was introduced. This processing operates directly on the Ambisonics signal [136, 230], and while defined in a closed mathematical form, leads to a trade-off between enhancement level and reproduction quality. Later research aimed to provide significant enhancement while perfectly preserving the desired spatial audio signal. Designed for Ambisonics signals, this aim is achieved by first estimating the DOA of the desired source, then estimating the source signal using high-directivity beamforming, and finally estimating the transfer function from the source signal to the Ambisonics signals. This process leads to a reconstruction of the desired Ambisonics signal with the full spatial information, while providing enhancement through the contribution of the beamforming [231]. Recently, this approach was also investigated for a wearable microphone array [143, 158]. An alternative approach, also aiming to achieve significant enhancement while preserving spatial information, employed masking in the time-frequency domain, applied directly to the Ambisonics signals or to the same signals spatially transformed by beamforming [232]. While high noise attenuation was achieved by masking in the transformed spatial domain, masking in the Ambisonics, or spherical harmonics domain, better preserved spatial information in the attenuated noise. Some of these approaches were generalized in a broad framework for signal enhancement [233], which incorporates source signal estimation under various sound field models in a way that preserves both the individual sources and the reverberant signal components, while minimizing the contribution of undesired noise.

While recent methods for the enhancement of spatial audio signals introduced a significant improvement compared to early methods, improved methods that provide superior performance in challenging environments with multiple speakers, reverberation and noise, may be highly desirable when considering realistic scenarios. In addition, many of the methods are designed for Ambisonics signals and spherical arrays, and so enhancement methods for more general array configurations may also be necessary.

5 Conclusion and outlook

This review paper presented an overview and recent developments in spatial audio signal processing, focusing on recorded sound and binaural reproduction. Significant progress has been made in the past decade, with the proposal of new methods and approaches, making a notable step towards providing high quality audio from recorded sound. Nevertheless, there are clear challenges ahead. These are outlined within the structure of the general framework presented in this paper.

- Acoustic scene – real-world scenes may be challenging, with several moving sources, reverberant environments, noise and interference. Most methods developed to date assume stationary sources, and so bridging the gap to handle several, and moving, sources in lively environments could be an important target for future research.

- Recording – spatial audio signals are recorded by microphone arrays, and with emerging applications such as smart homes and VR/AR, arrays may be of varying configurations (e.g., on a device), may be composed of only a few microphones, and may be dynamic in space (e.g., wearable arrays). With many of the current methods developed for spherical arrays, an important challenge is to extend emerging methods to work with general arrays and perform well even with moving arrays composed of only a few microphones. Wearable arrays may also introduce challenges with respect to the limited computation resources available, and latency constraints imposed by real-time reproduction and head-tracking, for example. Overcoming these challenges may open great opportunities for delivering affordable spatial audio for consumer devices.
- Processing – while signal processing has been the main topic of this paper and is incorporated in the points above as well, a main avenue of research that has been reviewed here is spatial audio signal processing based on learning from measured data. With deep-learning methods continuously developing, their incorporation in the challenging tasks outlined here could be of great benefit. Learning from measured data could also include parametric representation of sound fields based on microphone-array recordings, which have great potential for high performance with compact representations. Furthermore, emerging approaches for manipulating sound field information for translation and rotation, for example, by non-linear transformation of the directional space (i.e., warping), may lead to new possibilities and increased flexibility for VR/AR and other applications.
- Reproduction and perception – over headphones, and, in particular, using individualized HRTF, will probably be key to high quality spatial audio. The incorporation of individualized HRTF in state-of-the art algorithms is therefore essential. Furthermore, improved understanding of the relation between the processed audio signal and perception may be essential to ensure that important signal information is maintained or enhanced. Performance evaluations, currently mostly developed for listeners with head-tracking, should be extended to 6DoF motion. Also, mathematically formulated objectives, essential for machine- and deep-learning, that incorporate perceptual attributes, could be useful for developing data-based learning solutions that are perceptually motivated.

Conflict of interest

The authors declare no conflict of interest.

References

1. M.F. Davis: History of spatial coding. Journal of the Audio Engineering Society 51, 6 (2003) 554–569.
2. M. Vorländer: Past, present and future of dummy heads, in Proceedings of Acústica, Guimarães, Portugal, 2004, pp. 13–17.
3. D.R. Begault, E.M. Wenzel, M.R. Anderson: Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source. Journal of the Audio Engineering Society 49, 10 (2001) 904–916.
4. B. Xie: Head-related transfer function and virtual auditory display. 2nd ed., J. Ross Publishing, 2013.
5. M.A. Gerzon: Periphony: with-height sound reproduction. Journal of the Audio Engineering Society 21, 1 (February 1973) 2–10.
6. J.S. Bamford: An analysis of ambisonic sound systems of first and second order. PhD thesis, University of Waterloo, Ontario, Canada, 1995.
7. J. Daniel: Acoustic field representation, application to the transmission and the reproduction of complex sound environments in a multimedia context. PhD thesis, Université de Paris, Paris, France, 2000.
8. D.G. Malham, A. Myatt: 3-D sound spatialization using ambisonic techniques. Computer Music Journal 19, 4 (1995) 58–70.
9. M.A. Poletti: The design of encoding functions for stereophonic and polyphonic sound systems. Journal of the Audio Engineering Society 44, 11 (1996) 948–963.
10. T.D. Abhayapala, D.B. Ward: Theory and design of high order sound field microphones using spherical microphone array, in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Orlando, Florida, USA, 2002, pp. 1949–1952.
11. J. Meyer, G. Elko: A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield, in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Orlando, Florida, USA, 2002, pp. II-1781–II-1784.
12. M. Vorländer: Auralization: fundamentals of acoustics, modelling, simulation, algorithms and acoustic virtual reality. Springer, 2020.
13. J. Blauert, J. Braasch: The technology of binaural understanding. Springer, 2020.
14. H. Hacihabiboglu, E. De Sena, Z. Cvetkovic, J. Johnston, J. O. Smith III: Perceptual spatial audio recording, simulation, and rendering: an overview of spatial-audio techniques based on psychoacoustics. IEEE Signal Processing Magazine 34, 3 (2017) 36–54.
15. W. Zhang, P.N. Samarasinghe, H. Chen, T.D. Abhayapala: Surround by sound: a review of spatial audio recording and reproduction. Applied Sciences 7, 3 (2017) 532.
16. F. Zotter, M. Frank: Ambisonics: a practical 3D audio theory for recording, studio production, sound reinforcement, and virtual reality. Springer Nature, 2019.
17. D.P. Jarrett, E.A.P. Habets, P.A. Naylor: Theory and applications of spherical microphone array processing. Springer-Verlag, Berlin, 2017.
18. B. Rafaely, Fundamentals of spherical array processing. Springer-Verlag, Berlin, 2019.
19. J. Herre, J. Hilpert, A. Kuntz, J. Plogsties: MPEG-H 3D audio – the new standard for coding of immersive spatial audio. IEEE Journal of Selected Topics in Signal Processing 9, 5 (2015) 770–779.
20. V. Pulkki, S. Delikaris-Manias, A. Politis: Parametric time-frequency domain spatial audio. John Wiley & Sons, 2017.

21. K. Kowalczyk, O. Thiergart, M. Taseska, G. Del Galdo, V. Pulkki, E.P.A. Habets: Parametric Spatial Sound Processing: A flexible and efficient solution to sound scene acquisition, modification, and reproduction. IEEE Signal Processing Magazine 32, 2 (2015) 31–42.
22. V.R. Algazi, R.O. Duda: Headphone-based spatial sound. IEEE Signal Processing Magazine 28, 1 (2011) 33–42.
23. K. Sunder, J. He, E.L. Tan, W.-S. Gan: Natural sound rendering for headphones: integration of signal processing techniques. IEEE Signal Processing Magazine 32, 2 (2015) 100–113.
24. D.R. Begault, L.J. Trejo: 3-D sound for virtual reality and multimedia. NASA, Ames Research Center, Moffett Field, California, 2000, pp. 132–136.
25. P. Milgram, H. Takemura, A. Utsumi, F. Kishino: Augmented reality: a class of displays on the reality-virtuality continuum. Telemanipulator and Telepresence Technologies, International Society for Optics and Photonics, 1995, pp. 282–292.
26. V. Tourbabin, B. Rafaely: Analysis of distortion in audio signals introduced by microphone motion, in 2016 24th European Signal Processing Conference (EUSIPCO), Budapest, Hungary, 2016, pp. 998–1002.
27. A. Alexandridis, A. Griffin, A. Mouchtaris: Capturing and reproducing spatial audio based on a circular microphone array. Journal of Electrical and Computer Engineering 2013 (2013) 1–16.
28. I. Toshima, H. Uematsu, T. Hirahara: A steerable dummy head that tracks three-dimensional head movement: TeleHead. Acoustical Science and Technology 24 (09 2003) 327–329.
29. Zylia: Zylia ZM-1 microphone. Accessed on December 6, 2021. https://www.zylia.co/
30. T. Lokki: Subjective comparison of four concert halls based on binaural impulse responses. Acoustical Science and Technology 26, 2 (2005) 200–203.
31. T. Lokki, J. Pätynen, S. Tervo, S. Siltanen, L. Savioja: Engaging concert hall acoustics is made up of temporal envelope preserving reflections. The Journal of the Acoustical Society of America 129, 6 (2011) EL223–EL228.
32. O. Axelsson, M.E. Nilsson, B. Berglund: A principal components model of soundscape perception. The Journal of the Acoustical Society of America 128, 5 (2010) 2836–2846.
33. B. Boren, M. Musick, J. Grossman, A. Roginska: I hear NY4D: hybrid acoustic and augmented auditory display for urban soundscapes, in International Conference on Auditory Display, New York, NY, USA, 2014.
34. A. Leudar: An alternative approach to 3D audio recording and reproduction. Divergence Press 3, 1 (2014).
35. Eden Project: Rainforest at night: heart of darkness. Accessed on December 6, 2021. https://web.archive.org/web/20110719132826/http://www.edenproject.com/come-and-visit/whats-on/heart-of-darkness.php
36. H. Lee: Multichannel 3D microphone arrays: a review. Journal of the Audio Engineering Society 69, 1/2 (2021) 5–26.
37. B&K: Binaural microphone B&K type 4101-B. Accessed on December 6, 2021. https://www.bksv.com/en/transducers/acoustic/binaural/binaural-microphone?tab=overview
38. 3Dio: Free-space binaural microphone. Accessed on December 6, 2021. https://3diosound.com/products/free-space-binaural-microphone
39. Sennheiser: Sennheiser AMBEO VR mic. Accessed on December 6, 2021. https://en-us.sennheiser.com/microphone-3d-audio-ambeo-vr-mic
40. em32 Eigenmike array. mhAcoustics, 25 Summit Ave, Summit, NJ 07901, USA. Accessed on December 6, 2021. https://mhacoustics.com/products
41. R. Duraiswami, D. Zotkin, Z. Li, E. Grassi, N. Gumerov, L. Davis: High-order spatial audio capture and its binaural head-tracked playback over headphones with HRTF cues, in The 119th Convention of Audio Engineering Society, vol. 3, New York, NY, USA, 01 2005, pp. 1–16.
42. M. Noisternig, T. Musil, A. Sontacchi, R. Holdrich: 3D binaural sound reproduction using a virtual ambisonic approach, in IEEE International Symposium on Virtual Environments, Human-Computer Interfaces and Measurement Systems, 2003. VECIMS ‘03. 2003, IEEE,2003, pp. 174–178.
43. M. Fallahi, M. Hansen, S. Doclo, S. van de Par, D. Püschel, M. Blau: Evaluation of head-tracked binaural auralizations of speech signals generated with a virtual artificial head in anechoic and classroom environments. Acta Acustica 5 (2021) 30.
44. L. Madmoni, J. Donley, V. Tourbabin, B. Rafaely: Beamforming-based binaural reproduction by matching of binaural signals, in Audio Engineering Society Conference: International Conference on Audio for Virtual and Augmented Reality, 2020.
45. S. Sakamoto, J. Kodama, S. Hongo, T. Okamoto, Y. Iwaya, Y. Suzuki: A 3D sound-space recording system using spherical microphone array with 252ch microphones, in 20th International Congress on Acoustics 2010, ICA 2010 – Incorporating Proceedings of the 2010 Annual Conference of the Australian Acoustical Society, Sydney, Australia, 2010, pp. 3032–3035.
46. A. Roginska, P. Geluso: Immersive sound: the art and science of binaural and multi-channel audio, Taylor & Francis, 2017.
47. S. Werner, F. Klein, T. Mayenfels, K. Brandenburg: A summary on acoustic room divergence and its effect on externalization of auditory events, in 2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX), IEEE, 2016, pp. 1–6.
48. W.O. Brimijoin, A.W. Boyd, M.A. Akeroyd: The contribution of head movement to the externalization and internalization of sounds. PloS one 8, 12 (2013) e83068.
49. F.L. Wightman, D.J. Kistler: The importance of head movements for localizing virtual auditory display objects, in International Conference on Auditory Display, Georgia Institute of Technology, 1994.
50. M.-V. Laitinen, T. Pihlajamäki, S. Lösler, V. Pulkki: Influence of resolution of head tracking in synthesis of binaural audio, in Audio Engineering Society Convention 132, Audio Engineering Society, 2012.
51. P. Stitt, E. Hendrickx, J.-C. Messonnier, B. Katz: The influence of head tracking latency on binaural rendering in simple and complex sound scenes, in Audio Engineering Society Convention 140, Audio Engineering Society, 2016.
52. I. Engel, D.L. Alon, P.W. Robinson, R. Mehra: The effect of generic headphone compensation on binaural renderings, in Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio, Audio Engineering Society, 2019.
53. A. Lindau, F. Brinkmann: Perceptual evaluation of headphone compensation in binaural synthesis based on non-individual recordings. Journal of the Audio Engineering Society 60, 1/2 (2012) 54–62.
54. D. Pralong, S. Carlile: The role of individualized headphone calibration for the generation of high fidelity virtual auditory space, The Journal of the Acoustical Society of America 100, 6 (1996) 3785–3793.

55. F. Brinkmann, A. Lindau, S. Weinzierl: On the authenticity of individual dynamic binaural synthesis. The Journal of the Acoustical Society of America 142, 4 (2017) 1784–1795.
56. Z. Ben-Hur, D.L. Alon, R. Mehra, B. Rafaely: Binaural reproduction based on bilateral ambisonics and ear-aligned HRTFs. IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021) 901–913.
57. D. Griesinger: General overview of spatial impression, envelopment, localization, and externalization, in Audio Engineering Society Conference: 15th International Conference: Audio, Acoustics & Small Spaces, Copenhagen, Denmark, 1998.
58. T. Hidaka, T. Okano, L. Beranek: Interaural cross correlation (IACC) as a measure of spaciousness and envelopment in concert halls. The Journal of the Acoustical Society of America 92, 4 (1992) 2469–2469.
59. H. Lee: Capturing 360° audio using an equal segment microphone array (ESMA). Journal of the Audio Engineering Society 67, 1/2 (2019) 13–26.
60. T. Okano, L.L. Beranek, T. Hidaka: Relations among interaural cross-correlation coefficient (IACC_E), lateral fraction (LF_E), and apparent source width (ASW) in concert halls. The Journal of the Acoustical Society of America 104, 1 (1998) 255–265.
61. A. Lindau, V. Erbes, S. Lepa, H.-J. Maempel, F. Brinkman, S. Weinzierl: A spatial audio quality inventory (SAQI). Acta Acustica united with Acustica 100, 5 (2014) 984–994.
62. G. Lorho: Individual vocabulary profiling of spatial enhancement systems for stereo headphone reproduction, in Audio Engineering Society Convention 119, Audio Engineering Society, 2005.
63. C. Millns, H. Lee: An investigation into spatial attributes of 360° microphone techniques for virtual reality, in Audio Engineering Society Convention 144, Milan, Italy, 2018.
64. G. Reardon, A. Genovese, G. Zalles, P. Flanagan, A. Roginska: Evaluation of binaural renderers: multidimensional sound quality assessment, in Audio Engineering Society Conference: International Conference on Audio for Virtual and Augmented Reality, Redmons, WA, USA, 2018.
65. L.S.R. Simon, N. Zacharov, B.F.G. Katz: Perceptual attributes for the comparison of head-related transfer functions. The Journal of the Acoustical Society of America 140, 5 (2016) 3623–3632.
66. N. Zacharov, T. Pedersen, C. Pike: A common lexicon for spatial sound quality assessment – latest developments, in 2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX), Lisbon, Portugal, 2016, pp. 1–6.
67. A. Lindau, S. Weinzierl: Assessing the plausibility of virtual acoustic environments. Acta Acustica united with Acustica 98, 5 (2012) 804–810.
68. R.S. Pellegrini: Quality assessment of auditory virtual environments, in International Conference on Auditory Display, Helsinki, Finland, 2001.
69. J. Blauert: Spatial hearing: the psychophysics of human sound localization. MIT Press, 1997.
70. R. Baumgartner, P. Majdak, B. Laback: Modeling sound-source localization in sagittal planes for human listeners. The Journal of the Acoustical Society of America 136 (8 2014) 791–802.
71. V. Best, R. Baumgartner, M. Lavandier, P. Majdak, N. Kopčo: Sound externalization: a review of recent research. Trends in Hearing 24 (2020) 1–14.
72. S. Li, R. Baumgartner, J. Peissig: Modeling perceived externalization of a static, lateral sound image. Acta Acustica 4, 5 (2020) 21.
73. J. Reijniers, D. Vanderelst, C. Jin, S. Carlile, H. Peremans: An ideal-observer model of human sound localization. Biological Cybernetics 108, 2 (2014) 169–181.
74. R. Baumgartner, P. Majdak: Decision making in auditory externalization perception: model predictions for static conditions, Acta Acustica 5 (2021) 59.
75. F. Rumsey, S. Zieliński, R. Kassier: On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio quality. The Journal of the Acoustical Society of America 118, 2 (2005) 968–976.
76. I. Ananthabhotla, V.K. Ithapu, W.O. Brimijoin: A framework for designing head-related transfer function distance metrics that capture localization perception. JASA Express Letters 1, 4 (2021) 044401.
77. P. Majdak, R. Baumgartner: Computational models for listener-specific predictions of spatial audio quality, in EAA Spatial Audio Signal Processing Symposium, Paris, France, 2019, pp. 155–159.
78. T. Robotham, O.S. Rummukainen, J. Herre, E.A.P. Habets: Evaluation of binaural renderers in virtual reality environments: platform and examples, in Proc. of the 145th AES Convention, New York, NY, USA, 2018.
79. T. Robotham, O.S. Rummukainen, M. Kurz, M. Eckert, E. A.P. Habets: Comparing direct and indirect methods of audio quality evaluation in virtual reality scenes of varying complexity. IEEE Transactions on Visualization and Computer Graphics 28, 5 (2022) 2091–2101.
80. B.I. Băcilă, H. Lee: Listener-position and orientation dependency of auditory perception in an enclosed space: elicitation of salient attributes. Applied Sciences 11, 4 (2021) 1–24.
81. C. Schneiderwind, A. Neidhardt: Perceptual differences of position dependent room acoustics in a small conference room, in The International Symposium on Room Acoustics, Amsterdam, Netherlands, 2019.
82. V.R. Algazi, R.O. Duda, D.M. Thompson: Motion-tracked binaural sound. Journal of the Audio Engineering Society 52, 11 (2004) 1142–1156.
83. A. Lindau, S. Roos: Perceptual evaluation of discretization and interpolation for motion-tracked binaural (MTB-) recordings, in Proceedings of the 26th Tonmeistertagungm VDT International Convention, Leipzig, Germany, 2010, pp. 680–701.
84. S. Nagel, P. Jax: Dynamic binaural cue adaptation, in 2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC), IEEE, 2018, pp. 96–100.
85. P.G. Craven, M.A. Gerzon: Coincident microphone simulation covering three dimensional space and yielding various directional outputs, 1977. US Patent 4,042,779
86. P.B. Fellgett: Ambisonic reproduction of directionality in surround-sound systems. Nature 252, 5484 (1974) 534–538.
87. M.A. Gerzon: The design of precisely coincident microphone arrays for stereo and surround sound, in Audio Engineering Society Convention 50, Audio Engineering Society, 1975.
88. J.-M. Jot, V. Larcher, J.-M. Pernaux: A comparative study of 3-D audio encoding and rendering techniques, in Audio Engineering Society Conference: 16th International Conference: Spatial Sound Reproduction, Arktikum, Rovaniemi, Finland, 1999.
89. M.J. Evans, J.A.S. Angus, A.I. Tew: Analyzing head-related transfer function measurements using surface spherical harmonics. The Journal of the Acoustical Society of America 104, 4 (1998) 2400–2411.
90. B. Rafaely, A. Avni: Interaural cross correlation in a sound field represented by spherical harmonics. The Journal of the Acoustical Society of America 127, 2 (2010) 823–828.

91. A. Sontacchi, M. Noisternig, P. Majdak, R. Holdrich: An objective model of localisation in binaural sound reproduction systems, in Audio Engineering Society Conference: 21st International Conference: Architectural Acoustics and Sound Reinforcement, Audio Engineering Society, 2002.
92. Z. Ben-Hur, D. Alon, R. Mehra, B. Rafaely: Binaural reproduction using bilateral Ambisonics. Journal of the Audio Engineering Society, in AES International Conference on Audio for Virtual and Augmented Reality (AVAR), Redmond, WA, USA, August 2020, pp. 1–6.
93. A. Avni, J. Ahrens, M. Geier, S. Spors, H. Wierstorf, B. Rafaely: Spatial perception of sound fields recorded by spherical microphone arrays with varying spatial resolution. The Journal of the Acoustical Society of America 133, 5 (2013) 2711–2721.
94. T. Lübeck, H. Helmholz, J.M. Arend, C. Pörschmann, J. Ahrens: Perceptual evaluation of mitigation approaches of impairments due to spatial undersampling in binaural rendering of spherical microphone array data. Journal of the Audio Engineering Society 68, 6 (2020) 428–440.
95. M. Zaunschirm, C. Schörkhuber, R. Höldrich: Binaural rendering of ambisonic signals by head-related impulse response time alignment and a diffuseness constraint. The Journal of the Acoustical Society of America 143, 6 (2018) 3616–3627.
96. em32 Eigenmike microphone array release notes (v17. 0). mhAcoustics, 25 Summit Ave, Summit, NJ 07901, USA, 2013.
97. B. Rafaely: Plane-wave decomposition of the sound field on a sphere by spherical convolution. The Journal of the Acoustical Society of America 116, 4 (2004) 2149–2157.
98. B. Rafaely, B. Weiss, E. Bachmat: Spatial aliasing in spherical microphone arrays. IEEE Transactions on Signal Processing 55, 3 (2007) 1003–1010.
99. D.L. Alon, B. Rafaely: Beamforming with optimal aliasing cancellation in spherical microphone arrays. IEEE/ACM Transactions on Audio, Speech, and Language Processing 24, 1 (2016) 196–210.
100. D.L. Alon, B. Rafaely: Spatial decomposition by spherical array processing, in Parametric Time-Frequency Domain Spatial Audio, Chapter 2, V. Pulkki, S. Delikaris-Manias, A. Politis, Eds., Wiley.2017, pp. 25–47.
101. A. Wabnitz, N. Epain, C.T. Jin, A frequency-domain algorithm to upscale ambisonic sound scenes, in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, 2012, pp. 385–388.
102. A. Wabnitz, N. Epain, A. McEwan, C. Jin, Upscaling Ambisonic sound scenes using compressed sensing techniques, in IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 2011, pp. 1–4.
103. P.K.T. Wu, N. Epain, C. Jin: A super-resolution beamforming algorithm for spherical microphone arrays using a compressed sensing approach, in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, Canada, 2013, pp. 649–653.
104. N. Murata, S. Koyama, N. Takamune, H. Saruwatari: Sparse sound field decomposition with parametric dictionary learning for super-resolution recording and reproduction, in IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), Cancun, Mexico, 2015, pp. 69–72.
105. G. Routray, R.M. Hegde: Sparse plane-wave decomposition for upscaling ambisonic signals, in 2020 International Conference on Signal Processing and Communications (SPCOM), Bangalore, India, 2020, pp. 1–5.
106. G. Routray, S. Basu, P. Baldev, R.M. Hegde: Deep-sound field analysis for upscaling ambisonic signals, in EAA Spatial Audio Signal Processing Symposium, Paris, France, 2019, pp. 1–6.
107. L. Zhang, X. Wang, R. Hu, D. Li, W. Tu: Estimation of spherical harmonic coefficients in sound field recording using feed-forward neural networks. Multimedia Tools and Applications 80 (2021) 6187–6202.
108. L. Zhang, X. Wang, R. Hu, D. Li, W. Tu: Optimization of sound fields reproduction based higher-order ambisonics (HOA) using the generative adversarial network (GAN). Multimedia Tools and Applications 80, 2 (2021) 2205–2220.
109. Z. Ben-Hur, J. Sheaffer, B. Rafaely: Joint sampling theory and subjective investigation of plane-wave and spherical harmonics formulations for binaural reproduction. Applied Acoustics 134 (2018) 138–144.
110. Z. Ben-Hur, F. Brinkmann, J. Sheaffer, S. Weinzierl, B. Rafaely: Spectral equalization in binaural signals represented by order-truncated spherical harmonics. The Journal of the Acoustical Society of America 141, 6 (2017) 4087–4096.
111. C. Hold, H. Gamper, V. Pulkki, N. Raghuvanshi, I.J. Tashev: Improving binaural ambisonics decoding by spherical harmonics domain tapering and coloration compensation, in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019, pp. 261–265.
112. C. Schörkhuber, M. Zaunschirm, R. Höldrich: Binaural rendering of ambisonic signals via magnitude least squares, in Fortschritte der Akustik (DAGA), München, Germany, 2018, pp. 339–342.
113. F. Brinkmann, S. Weinzierl: Comparison of head-related transfer functions pre-processing techniques for spherical harmonics decomposition, in Audio Engineering Society Conference: International Conference on Audio for Virtual and Augmented Reality, Redmons, WA, USA, 2018.
114. L. Birnie, T. Abhayapala, P. Samarasinghe, V. Tourbabin: Sound field translation methods for binaural reproduction, in IX-Degrees-of-Freedom Binaural IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), IEEE, 2019, pp. 140–144.
115. H. Lee, F. Rumsey: Level and time panning of phantom images for musical sources. Journal of the Audio Engineering Society 61, 12 (2013) 978–988.
116. M. Williams, G. Le Du: Microphone array analysis for multichannel sound recording, in Audio Engineering Society Convention 107, New York, NY, USA, 1999.
117. H. Wittek, G. Theile: The recording angle – based on localisation curves, in Audio Engineering Society Convention 112, Munich, Germany, 2002.
118. F. Zotter, M. Frank: Efficient phantom source widening. Archives of Acoustics 38, 1 (2013) 27–37.
119. K. Hamasaki, K. Hiyama: Reproducing spatial impression with multichannel audio, in Audio Engineering Society Conference: 24th International Conference: Multichannel Audio, The New Reality, Banff, Alberta, Canada, 2003.
120. F. Rumsey: Spatial audio, Focal Press, 2001.
121. M. Kuster: Spatial correlation and coherence in reverberant acoustic fields: extension to microphones with arbitrary first-order directivity. The Journal of the Acoustical Society of America 123, 1 (2008) 154–162.
122. D. Griesinger: Reproducing low frequency spaciousness and envelopment in listening rooms, in Audio Engineering Society Convention 145, New York, NY, USA, 2018.
123. C. Gribben, H. Lee: A comparison between horizontal and vertical interchannel decorrelation. Applied Sciences 7, 11 (2017) 1–21.

124. C. Gribben, H. Lee: The frequency and loudspeaker-azimuth dependencies of vertical interchannel decorrelation on the vertical spread of an auditory image. Journal of the Audio Engineering Society 66, 7/8 (2018) 537–555.
125. H. Lee, C. Gribben: Effect of vertical microphone layer spacing for a 3D microphone array. Journal of the Audio Engineering Society 62, 12 (2014) 870–884.
126. H. Wittek, G. Theile: Development and application of a stereophonic multichannel recording technique for 3D audio and VR, in 143rd International Convention of the Audio Engineering Society, Audio Engineering Society, 2017.
127. H. Lee, M. Frank, F. Zotter: Spatial and timbral fidelities of binaural ambisonics decoders for main microphone array recordings, in Audio Engineering Society Conference: International Conference on Immersive and Interactive Audio, York, UK, 2019.
128. A. McKeag, D.S. McGrath: Sound field format to binaural decoder with head tracking, in 6th Austrailian Regional Convention of the AES, Audio Engineering Society, 1996.
129. A.M. O'Donovan, D.N. Zotkin, R. Duraiswami: Spherical microphone array based immersive audio scene rendering, in International Conference on Auditory Display,2008.
130. J. Jiang, B. Xie, H. Mai: The number of virtual loudspeakers and the error for spherical microphone array recording and binaural rendering, in Audio Engineering Society Conference: International Conference on Spatial Reproduction-Aesthetics and Science, Tokyo, Japan, 2018.
131. H.L. Van Trees: Optimum array processing. John Wiley & Sons, 2002.
132. W. Song, W. Ellermeier, J. Hald: Binaural auralization based on spherical-harmonics beamforming. The Journal of the Acoustical Society of America 123, 5 (2008) 3159–3159.
133. W. Song, W. Ellermeier, J. Hald: Psychoacoustic evaluation of multichannel reproduced sounds using binaural synthesis and spherical beamforming. The Journal of the Acoustical Society of America 130, 4 (2011) 2063–2075.
134. W. Song, W. Ellermeier, J. Hald: Using beamforming and binaural synthesis for the psychoacoustical evaluation of target sources in noise. The Journal of the Acoustical Society of America 123, 2 (2008) 910–924.
135. S. Spors, H. Wierstorf, M. Geier: Comparison of modal versus delay-and-sum beamforming in the context of data-based binaural synthesis, in Audio Engineering Society Convention 132, Budapest, Hungary, April 2012.
136. M. Jeffet, N.R. Shabtai, B. Rafaely: Theory and perceptual evaluation of the binaural reproduction and beamforming tradeoff in the generalized spherical array beamformer. IEEE/ACM Transactions on Audio, Speech, and Language Processing 24, 4 (2016) 708–718.
137. N.R. Shabtai, B. Rafaely: Binaural sound reproduction beamforming using spherical microphone arrays, in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, Canada, 2013, pp. 101–105.
138. N.R. Shabtai, B. Rafaely: Spherical array beamforming for binaural sound reproduction, in IEEE Convention of Electrical and Electronics Engineers in Israel, Eilat, Israel, 2012, pp. 1–5.
139. N.R. Shabtai: Optimization of the directivity in binaural sound reproduction beamforming. The Journal of the Acoustical Society of America 138, 5 (2015) 3118–3128.
140. E. Hadad, D. Marquardt, S. Doclo, S. Gannot: Theoretical analysis of binaural transfer function MVDR beamformers with interference cue preservation constraints. IEEE/ACM Transactions on Audio, Speech, and Language Processing 23, 12 (2015) 2449–2464.
141. E. Hadad, S. Doclo, S. Gannot: The binaural LCMV beamformer and its performance analysis. IEEE/ACM Transactions on Audio, Speech, and Language Processing 24, 3 (2016) 543–558.
142. P. Calamia, S. Davis, C. Smalt, C. Weston: A conformal, helmet-mounted microphone array for auditory situational awareness and hearing protection, in IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 2017, pp. 96–100.
143. H. Beit-On, M. Lugasi, L. Madmoni, A. Menon, A. Kumar, J. Donley, V. Tourbabin, B. Rafaely: Audio signal processing for telepresence based on wearable array in noisy and dynamic scenes, in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Singapore, 2022, accepted for publication.
144. M. Blau, A. Budnik, M. Fallahi, H. Steffens, S.D. Ewert, S. Van de Par: Toward realistic binaural auralizations–perceptual comparison between measurement and simulation-based auralizations and the real room for a classroom scenario. Acta Acustica 5 (2021) 8.
145. I. Ifergan, B. Rafaely: On the selection of the number of beamformers in beamforming-based binaural reproduction. EURASIP Journal on Audio, Speech and Music Processing 6 (2022) 1–17.
146. D. Marelli, R. Baumgartner, P. Majdak: Efficient approximation of head-related transfer functions in subbands for accurate sound localization. IEEE/ACM Transactions on Audio, Speech, and Language Processing 23, 7 (2015) 1130–1143.
147. V. Pulkki: Spatial sound reproduction with directional audio coding. Journal of the Audio Engineering Society 55, 6 (2007) 503–516.
148. M.M. Goodwin, J.-M. Jot: Primary-ambient signal decomposition and vector-based localization for spatial audio coding and enhancement, in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Honolulu, Hawaii, USA, 2007, pp. I-9–I-12.
149. N. Barrett, S. Berge: A new method for B-format to binaural transcoding, in Audio Engineering Society Conference: 40th International Conference: Spatial Audio: Sense the Sound of Space, Audio Engineering Society, 2010.
150. S. Berge, B. Allmenndigitale, N. Barrett: High angular resolution planewave expansion, in Proceedings of the 2nd International Symposium on Ambisonics and Spherical Acoustics, Paris, France, 2010.
151. O. Thiergart, E.A.P. Habets: Parametric sound acquisition using a multi-wave signal model and spatial filters, in Parametric Time-Frequency Domain Spatial Audio, V. Pulkki, S. Delikaris-Manias, A. Politis, Eds., John Wiley & Sons. 2017.
152. O. Thiergart, M. Taseska, E.A.P. Habets: An informed parametric spatial filter based on instantaneous direction-of-arrival estimates. IEEE/ACM Transactions on Audio, Speech, and Language Processing 22, 12 (2014) 2182–2196.
153. C.T. Jin, Y. Shiduo, F. Antonacci, A. Sarti: Perspectives on microphone array processing including sparse recovery, ray space analysis, and neural networks. Acoustical Science and Technology 41, 1 (2020) 308–317.
154. A. Politis, J. Vilkamo, V. Pulkki: Sector-based parametric sound field reproduction in the spherical harmonic domain. IEEE Journal of Selected Topics in Signal Processing 9, 5 (2015) 852–866.
155. V. Pulkki, A. Politis, G. Del Galdo, A. Kuntz: Parametric spatial audio reproduction with higher-order B-format microphone input, in Audio Engineering Society Convention 134, Audio Engineering Society, 2013.
156. A. Politis, S. Tervo, V. Pulkki: Compass: Coding and multidirectional parameterization of ambisonic sound scenes, in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 6802–6806.

157. L. McCormack, A. Politis, R. Gonzalez, T. Lokki, V. Pulkki: Parametric ambisonic encoding of arbitrary microphone arrays. IEEE/ACM Transactions on Audio, Speech, and Language Processing 30 (2022) 2062–2075.
158. J. Fernandez, L. McCormack, P. Hyvärinen, A. Politis, V. Pulkki: Enhancing binaural rendering of head-worn microphone arrays through the use of adaptive spatial covariance matching. The Journal of the Acoustical Society of America 151, 4 (2022) 2624–2635.
159. L. McCormack, A. Politis, V. Pulkki: Rendering of source spread for arbitrary playback setups based on spatial covariance matching, in IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, USA, 2021.
160. J. Daniel, S. Kitić: Echo-enabled direction-of-arrival and range estimation of a mobile source in Ambisonic domain, 2022. arXiv preprint arXiv:2203.05265
161. S. Kitić, J. Daniel: Generalized time domain velocity vector, in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 2022, pp. 936–940.
162. T. Shlomo, B. Rafaely: Blind amplitude estimation of early room reflections using alternating least squares, in ICASSP 2021 – 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, Canada, 2021, pp. 476–480.
163. T. Shlomo, B. Rafaely: Blind localization of early room reflections using phase aligned spatial correlation. IEEE Transactions on Signal Processing 69 (2021) 1213–1225.
164. IEEE AASP challenge on detection and classification of acoustic scenes and events (DCASE). Accessed on December 6, 2021. http://dcase.community/challenge2021/
165. A. Mesaros, T. Heittola, T. Virtanen: A multi-device dataset for urban acoustic scene classification, in Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018), Surrey, UK, 2018, pp. 9–13.
166. A. Politis, S. Adavanne, T. Virtanen: A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection, in Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE2020), 2020.
167. P.-A. Grumiaux: Deep learning for speaker counting and localization with Ambisonics signals. PhD thesis, Université Grenoble Alpes (UGA), 2021.
168. J. Eaton, N.D. Gaubitch, A.H. Moore, P.A. Naylor: Estimation of room acoustic parameters: the ACE challenge. IEEE/ACM Transactions on Audio, Speech, and Language Processing 24, 10 (2016) 1681–1693.
169. H. Gamper, I.J. Tashev: Blind reverberation time estimation using a convolutional neural network, in 2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC), IEEE, 2018, pp. 136–140.
170. P. Götz, C. Tuna, A. Walther, E.A.P. Habets: Blind reverberation time estimation in dynamic acoustic conditions, in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 2022.
171. S. Deng, W. Mack, E.A.P. Habets: Online blind reverberation time estimation using CRNNs, in INTERSPEECH, Incheon, Korea, 2020, pp. 5061–5065.
172. S. Duangpummet, J. Karnjana, W. Kongprawechnon, M. Unoki: Blind estimation of room acoustic parameters and speech transmission index using MTF-based CNNs, in The European Signal Processing Conference (EUSIPCO), Dublin, Ireland, 2021, pp. 181–185, abs/2103.07904
173. D. Looney, N.D. Gaubitch: Joint estimation of acoustic parameters from single-microphone speech observations, in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 431–435.
174. P. Morgado, N. Vasconcelos, T. Langlois, O. Wang: Self-supervised generation of spatial audio for 360 video, 2018. arXiv preprint arXiv:1809.02587
175. A. Richard, D. Markovic, I.D. Gebru, S. Krenn, G.A. Butler, F. Torre, Y. Sheikh: Neural synthesis of binaural speech from mono audio, in International Conference on Learning Representations, 2021.
176. M. Cobos, J. Ahrens, K. Kowalczyk, A. Politis: An overview of machine learning and other data-based methods for spatial audio capture, processing, and reproduction. EURASIP Journal on Audio, Speech, and Music Processing 2022, 1 (2022) 1–21.
177. HEAR360: 8Ball microphone. Accessed on December 6, 2021. https://8ballmicrophones.com
178. 3DOI: Omni binaural microphone. Accessed on December 6, 2021. https://3diosound.com/products/omni-binaural-microphone
179. M. Noisternig, A. Sontacchi, T. Musil, R. Holdrich: A 3D ambisonic based binaural sound reproduction system, in Audio Engineering Society Conference: 24th International Conference: Multichannel Audio, The New Reality, 2003.
180. L.S. Davis, R. Duraiswami, E. Grassi, N.A. Gumerov, Z. Li, D.N. Zotkin: High order spatial audio capture and its binaural head-tracked playback over headphones with HRTF cues, in Audio Engineering Society Convention 119, Audio Engineering Society, 2005.
181. C.H. Choi, J. Ivanic, M.S. Gordon, K. Ruedenberg: Rapid and stable determination of rotation matrices between spherical harmonics by direct recursion. The Journal of Chemical Physics 111, 19 (1999) 8825–8831.
182. N.A. Gumerov, R. Duraiswami: Fast multipole methods for the helmholtz equation in three dimensions. Elsevier, 2005.
183. P.J. Kostelec, D.N. Rockmore: FFTs on the rotation group. Journal of Fourier Analysis and Applications 14, 2 (2008) 145–179.
184. D. Pinchon, P.E. Hoggan: Rotation matrices for real spherical harmonics: general rotations of atomic orbitals in space-fixed axes. Journal of Physics A: Mathematical and Theoretical 40, 7 (2007) 1597.
185. B. Rafaely, M. Kleider: Spherical microphone array beam steering using Wigner-D weighting. IEEE Signal Processing Letters 15 (2008) 417–420.
186. F. Zotter: Analysis and synthesis of sound-radiation with spherical arrays. PhD thesis, University of Music and Performing Arts, Vienna, Austria, 2009.
187. J. Ahrens, H. Helmholz, D.L. Alon, S.V.A. Garí: A head-mounted microphone array for binaural rendering, in 2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA), IEEE, 2021, pp. 1–7.
188. J. Ahrens, H. Helmholz, D.L. Alon, S.V.A. Garí: Spherical harmonic decomposition of a sound field based on microphones around the circumference of a human head, in IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), IEEE, 2021, pp. 231–235.
189. L. Madmoni, J. Donley, V. Tourbabin, B. Rafaely: Binaural reproduction from microphone array signals incorporating head-tracking, in 2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA), IEEE, 2021, pp. 1–5.
190. D. Rivas Méndez, C. Armstrong, J. Stubbs, M. Stiles, G. Kearney: Practical recording techniques for music production with six-degrees of freedom virtual reality, in Audio Engineering Society Convention 145, Audio Engineering Society, 2018.

191. J. Daniel: Spatial sound encoding including near field effect: Introducing distance coding filters and a viable, new ambisonic format, in Audio Engineering Society Conference: 23rd International Conference: Signal Processing in Audio Recording and Reproduction, Copenhagen, Denmark, 2003.
192. E. Stein, M.M. Goodwin: Ambisonics depth extensions for six degrees of freedom, in Audio Engineering Society Conference: International Conference on Headphone Technology, San Francisco, CA, USA, 2019.
193. F. Zotter, M. Frank, C. Schörkhuber, R. Höldrich: Signal-independent approach to variable-perspective (6DoF) audio rendering from simultaneous surround recordings taken at multiple perspectives, in Fortschritte der Akustik (DAGA), Hannover, Germany, 2020.
194. E. Bates, H. O'Dwyer, K.-P. Flachsbarth, F.M. Boland: A recording technique for 6 degrees of freedom VR, in Audio Engineering Society Convention 144, Audio Engineering Society, 2018.
195. E. Fernandez-Grande: Sound field reconstruction using a spherical microphone array. The Journal of the Acoustical Society of America 139, 3 (2016) 1168–1178.
196. T. Pihlajamaki, V. Pulkki: Synthesis of complex sound scenes with transformation of recorded spatial sound in virtual reality. Journal of the Audio Engineering Society 63, 7/8 (2015) 542–551.
197. A. Plinge, S.J. Schlecht, O. Thiergart, T. Robotham, O. Rummukainen, E.A.P. Habets: Six-degrees-of-freedom binaural audio reproduction of first-order Ambisonics with distance information, in Audio Engineering Society Conference: International Conference on Audio for Virtual and Augmented Reality, 2018.
198. K. Wakayama, J. Trevino, H. Takada, S. Sakamoto, Y. Suzuki: Extended sound field recording using position information of directional sound sources, in IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), IEEE, 2017, pp. 185–189.
199. A. Allen, B. Kleijn: Ambisonics soundfield navigation using directional decomposition and path distance estimation, in International Conference on Spatial Audio, Graz, Austria, 2017.
200. M. Kentgens, A. Behler, P. Jax, Translation of a higher order Ambisonics sound scene based on parametric decomposition, in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 151–155.
201. F. Schultz, S. Spors: Data-based binaural synthesis including rotational and translatory head-movements, in Audio Engineering Society Conference: 52nd International Conference: Sound Field Control-Engineering and Perception, Guildford, UK, 2013.
202. Y. Wang, K. Chen: Translations of spherical harmonics expansion coefficients for a sound field using plane wave expansions. The Journal of the Acoustical Society of America 143, 6 (2018) 3474–3478.
203. L. Birnie, T. Abhayapala, V. Tourbabin, P. Samarasinghe: Mixed source sound field translation for virtual binaural application with perceptual validation. IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021) 1188–1203.
204. J.G. Tylka, E. Choueiri: Comparison of techniques for binaural navigation of higher-order ambisonic soundfields, in Audio Engineering Society Convention 139, Audio Engineering Society, 2015.
205. J.G. Tylka, E.Y. Choueiri: Performance of linear extrapolation methods for virtual sound field navigation. Journal of the Audio Engineering Society 68, 3 (2020) 138–156.
206. M. Kentgens, P. Jax: Ambient-aware sound field translation using optimal spatial filtering, in IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), IEEE, 2021, pp. 236–240.
207. M. Kentgens, S. Al Hares, P. Jax: On the upscaling of higher-order Ambisonics signals for sound field translation, in 2021 29th European Signal Processing Conference (EUSIPCO), IEEE, 2021, pp. 81–85.
208. A. Brutti, M. Omologo, P. Svaizer: Localization of multiple speakers based on a two step acoustic map analysis, in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Las Vegas, NV, USA, 2008, pp. 4349–4352.
209. A. Brutti, M. Omologo, P. Svaizer: Multiple source localization based on acoustic map de-emphasis. EURASIP Journal on Audio, Speech, and Music Processing 2010 (2010) 1–17.
210. G. Del Galdo, O. Thiergart, T. Weller, E.A.P. Habets: Generating virtual microphone signals using geometrical information gathered by distributed arrays, in 2011 Joint Workshop on Hands-free Speech Communication and Microphone Arrays, IEEE, 2011, pp. 185–190.
211. O. Thiergart, G. Del Galdo, M. Taseska, E.A.P. Habets: Geometry-based spatial sound acquisition using distributed microphone arrays. IEEE Transactions on Audio, Speech, and Language Processing 21, 12 (2013) 2583–2594.
212. X. Zheng: Soundfield navigation: separation, compression and transmission. PhD thesis, University of Wollongong, Wollongong, Australia, 2013.
213. J.G. Tylka, E. Choueiri: Soundfield navigation using an array of higher-order Ambisonics microphones, in Audio Engineering Society Conference: International Conference on Audio for Virtual and Augmented Reality, Los Angeles, CA, USA, 2016.
214. J.G. Tylka, E.Y. Choueiri: Domains of practical applicability for parametric interpolation methods for virtual sound field navigation. Journal of the Audio Engineering Society 67, 11 (2019) 882–893.
215. J.G. Tylka: Virtual navigation of Ambisonics-encoded sound fields containing near-field sources. PhD thesis, Princeton University, Princeton, USA, 2019.
216. M.F. Fallon, S.J. Godsill: Acoustic source localization and tracking of a time-varying number of speakers. IEEE Transactions on Audio, Speech, and Language Processing 20, 4 (2011) 1409–1415.
217. S. Kitić, A. Guérin: Tramp: tracking by a real-time ambisonic-based particle filter, in Proceedings of LOCATA Challenge Workshop – a satellite event of IWAENC 2018, Tokyo, Japan, 2018.
218. J.-M. Valin, F. Michaud, J. Rouat: Robust 3D localization and tracking of sound sources using beamforming and particle filtering, in IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), vol. 4, Toulouse, France, 2006, IV–841–IV–844.
219. J.-M. Valin, F. Michaud, J. Rouat: Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering. Robotics and Autonomous Systems 55, 3 (2007) 216–228.
220. D.B. Ward, E.A. Lehmann, R.C. Williamson: Particle filtering algorithms for tracking an acoustic source in a reverberant environment. IEEE Transactions on Speech and Audio Processing 11, 6 (2003) 826–836.
221. N. Mariette, B.F.G. Katz, K. Boussetta, O. Guillerminet: Sounddelta: a study of audio augmented reality using wifi-distributed ambisonic cell rendering, in Audio Engineering Society Convention 128, Audio Engineering Society, 2010.

222. E. Patricio, A. Ruminski, A. Kuklasinski, L. Januszkiewicz, T. Zernicki: Toward six degrees of freedom audio recording and playback using multiple Ambisonics sound fields, in Audio Engineering Society Convention 146, Audio Engineering Society, 2019.
223. C. Schörkhuber, R. Höldrich, F. Zotter: Triplet-based variable-perspective (6DoF) audio rendering from simultaneous surround recordings taken at multiple perspectives, in Fortschritte der Akustik (DAGA), vol. 4, Hannover, Germany, 2020.
224. P. Grosche, F. Zotter, C. Schörkhuber, M. Frank, R. Höldrich: Method and apparatus for acoustic scene playback, 2020. US Patent 10,785,588.
225. M. Blochberger, F. Zotter: Particle-filter tracking of sounds for frequency-independent 3D audio rendering from distributed B-format recordings. Acta Acustica 5 (2021) 20.
226. L. McCormack, A. Politis, T. McKenzie, C. Hold, V. Pulkki: Object-based six-degrees-of-freedom rendering of sound scenes captured with multiple Ambisonic receivers. Journal of the Audio Engineering Society 70, 5 (2022) 355–372.
227. E. Erdem, O. Olgun, H. Hacihabiboğlu: Internal time delay calibration of rigid spherical microphone arrays for multi-perspective 6DoF audio recordings, in IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), IEEE, 2021, pp. 241–245.
228. O. Olgun, E. Erdem, H. Hachabiboğlu: Rotation calibration of rigid spherical microphone arrays for multi-perspective 6DoF audio recordings, in 2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA), IEEE, 2021, pp. 1–7.
229. A.H. Moore, L. Lightburn, W. Xue, P.A. Naylor, M. Brookes: Binaural mask-informed speech enhancement for hearing aids with head tracking, in 2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC), Tokyo, Japan, 2018, pp. 461–465.
230. N.R. Shabtai, B. Rafaely: Generalized spherical array beamforming for binaural speech reproduction. IEEE/ACM Transactions on Audio, Speech, and Language Processing 22, 1 (2013) 238–247.
231. C. Borrelli, A. Canclini, F. Antonacci, A. Sarti, S. Tubaro: A denoising methodology for higher order Ambisonics recordings, in 2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC), IEEE, 2018, pp. 451–455.
232. M. Lugasi, B. Rafaely: Speech enhancement using masking for binaural reproduction of Ambisonics signals. IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2020) 1767–1777.
233. A. Herzog, E.A.P. Habets: Direction and reverberation preserving noise reduction of ambisonics signals. IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2020) 2461–2475.

Cite this article as: Rafaely B. Tourbabin V. Habets E. Ben-Hur Z. Lee H, et al. 2022. Spatial audio signal processing for binaural reproduction of recorded acoustic scenes – review and challenges. Acta Acustica, **6**, 47.