CHALMERS
UNIVERSITY OF TECHNOLOGY

An overview of machine learning and other data-based methods for spatial audio capture, processing, and reproduction

Cobos *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2022) 2022:10
https://doi.org/10.1186/s13636-022-00242-x

EURASIP Journal on Audio, Speech, and Music Processing

REVIEW Open Access

An overview of machine learning and other data-based methods for spatial audio capture, processing, and reproduction

Maximo Cobos[1*†], Jens Ahrens[2*†], Konrad Kowalczyk[3*†] and Archontis Politis[4*†]

Abstract

The domain of spatial audio comprises methods for capturing, processing, and reproducing audio content that contains spatial information. Data-based methods are those that operate directly on the spatial information carried by audio signals. This is in contrast to model-based methods, which impose spatial information from, for example, metadata like the intended position of a source onto signals that are otherwise free of spatial information. Signal processing has traditionally been at the core of spatial audio systems, and it continues to play a very important role. The irruption of deep learning in many closely related fields has put the focus on the potential of learning-based approaches for the development of data-based spatial audio applications. This article reviews the most important application domains of data-based spatial audio including well-established methods that employ conventional signal processing while paying special attention to the most recent achievements that make use of machine learning. Our review is organized based on the topology of the spatial audio pipeline that consist in capture, processing/manipulation, and reproduction. The literature on the three stages of the pipeline is discussed, as well as on the spatial audio representations that are used to transmit the content between them, highlighting the key references and elaborating on the underlying concepts. We reflect on the literature based on a juxtaposition of the prerequisites that made machine learning successful in domains other than spatial audio with those that are found in the domain of spatial audio as of today. Based on this, we identify routes that may facilitate future advancement.

Keywords: Spatial audio, Machine learning, Deep learning, Array processing, Ambisonics, Virtual reality, Binaural audio, Audio coding, Scene analysis

1 Introduction

Interest in immersive communication technologies has been growing over the last two decades due to the emergence of today's multimedia applications. Gaming, virtual and augmented reality (VR and AR), teleconferencing, and entertainment applications have taken a big step forward due to the spread of mobile multimedia and ubiquitous computing [1]. Spatial audio research is at the center of the new developments in 3D immersive user experiences, providing the core technologies for spatial sound capture, processing, and reproduction [2].

Spatial audio is an interdisciplinary field of research that brings together experts from audio engineering, acoustics, computer science, applied psychoacoustics, and other domains. The aim of spatial audio is to recreate an acoustic environment or synthesize a new one by using a proper combination of sound recording, processing, and reproduction techniques. Within such an objective, it is not only important to preserve the fidelity of the audio content

†Maximo Cobos, Jens Ahrens, Konrad Kowalczyk and Archontis Politis contributed equally to this work.
*Correspondence: maximo.cobos@uv.es; jens.ahrens@chalmers.se; konrad.kowalczyk@agh.edu.pl; archontis.politis@tuni.fi
[1]Computer Science Department, Universitat de València, 46100 Burjassot, Valencia, Spain
[2]Division of Applied Acoustics, Chalmers University of Technology, 412 96 Gothenburg, Sweden
[3]Institute of Electronics, AGH University of Science and Technology, 30-059 Krakow, Poland
[4]Department of Information Technology and Communication Sciences, Tampere University, Tampere FI-33720, Finland

but also the spatial attributes of the sound scene resulting from the actual locations of the sound sources and the properties of the acoustic environment [3–5].

In general, spatial audio methods and techniques may be broadly described as illustrated in Fig. 1. A complete spatial audio pipeline generally comprises a *capture* stage, a *processing* stage in which the spatial information in the captured sound scene is modified or in which given information is extracted from the sound scene that cannot be measured directly, and finally a *reproduction* stage in which the (potentially manipulated) sound scene is auralized. A more detailed overview of these stages, which are central to the organization of this overview, is provided in Section 1.1.

While a hard classification of spatial audio methods can be difficult to establish, many of them can be broadly categorized as model-based or data-based [6]. Traditionally, model-based methods[1] compose sound scenes from individual virtual sound sources that are described analytically by mathematical or physical models and driven by a set of audio input signals. Wave field synthesis (WFS) [8], stereophonic amplitude panning, and vector base amplitude panning (VBAP) [9] are all model-based methods.

In contrast, data-based spatial audio methods[2] employ sound scene representations in which the spatial information is encoded in the audio signals. The spatial information can originate from array recordings, acoustic measurements, or from simulations. Section 1.2 elaborates more on the fundamental differences existing between model-based and data-based spatial audio.

We also mention the concept of *object-based* audio representation here [12]. This concept is very similar to model-based representation in that a spatial scene is represented by its components. Audio objects can be more abstract than the objects in a model-based representation. Data-based reverberation, for example, can be an audio object in an otherwise model-based scene.

Acoustics and signal processing have been traditionally highly intertwined in the development of spatial audio techniques [13]. Signal processing algorithms for ambience extraction, personalization of head-related transfer functions (HRTFs), audio up-mixing, and sound field rendering have been available for several decades and are still finding application in current multimedia systems. Most traditional methods in spatial audio have been designed from a pure signal processing perspective. The irruption of deep learning (DL) [14] in the recent years is starting to create a turning point in many areas of digital signal processing, and consequently, spatial audio is also starting to feel the impact of machine learning (ML) in general, and deep neural networks (DNNs) in particular.

ML comprises a learning process that enables ML models to recognize patterns of interest in the data on which the systems are trained and to apply that knowledge to detect or generate similar patterns on new, unseen data. We highlight at this point that the term *data* when used in an ML context can refer to any type of data, be it audio signals, digital images, financial transactions, or others. The *data* in data-based spatial audio, on the other hand, are always spatial information that is encoded in the signals. Throughout this article, the term will primarily refer to multichannel audio data with spatial cues encoded as inter-channel dependencies. As will be shown by numerous examples of audio applications, such data are suitable for ML.

Undoubtedly, the popularity of DL in image processing, computer vision, and natural language processing has led to significant impact in fields closely related to spatial audio, including speech enhancement or music information retrieval [15, 16]. While ML algorithms have already positioned themselves at the top of the state of the art within the aforementioned fields, their use in immersive spatial audio is only emerging, as it will be illustrated throughout this review.

This article provides an overview of data-based spatial audio methods and establishes a topology of the concepts that have been employed. The scope in which data-based methods have been utilized in spatial audio capture, processing, and reproduction is broad, and the potential of DL has been in the focus particularly in the recent couple of years. We complement our review with the relevant works on signal-processing-based (i.e., non-ML-based) methods that are sometimes alternatives to the ML-based methods and complementary at other times.

With the aim of elaborating better on the role of data-based methods within the general spatial audio pipeline, the remaining sections of this introduction are devoted to introducing these two important aspects, and it wraps up with formalizing the article scope.

1.1 The spatial audio pipeline

This overview is organized based on the topography of the spatial audio pipeline. The spatial audio pipeline (Fig. 1) starts with a given representation of a sound field including spatial information. Usually, spatial sound scenes are captured using an array of microphones with a given geometry (cf. Fig. 2 for examples). The microphone output signals together with the microphones' positions and their directivity already constitute a representation of the sound scene. In some setups, the microphone signals are combined by suitable mathematical operations to obtain

[1] These are not to be confused with model-based signal processing [7].
[2] The term *data-based* in this context was originally introduced in [10, 11] for audio reproduction based on databases of room impulse responses (that encoded the spatial information). It has been subsequently used for all types of sound scene representations in which the audio signals encode the spatial information [12].

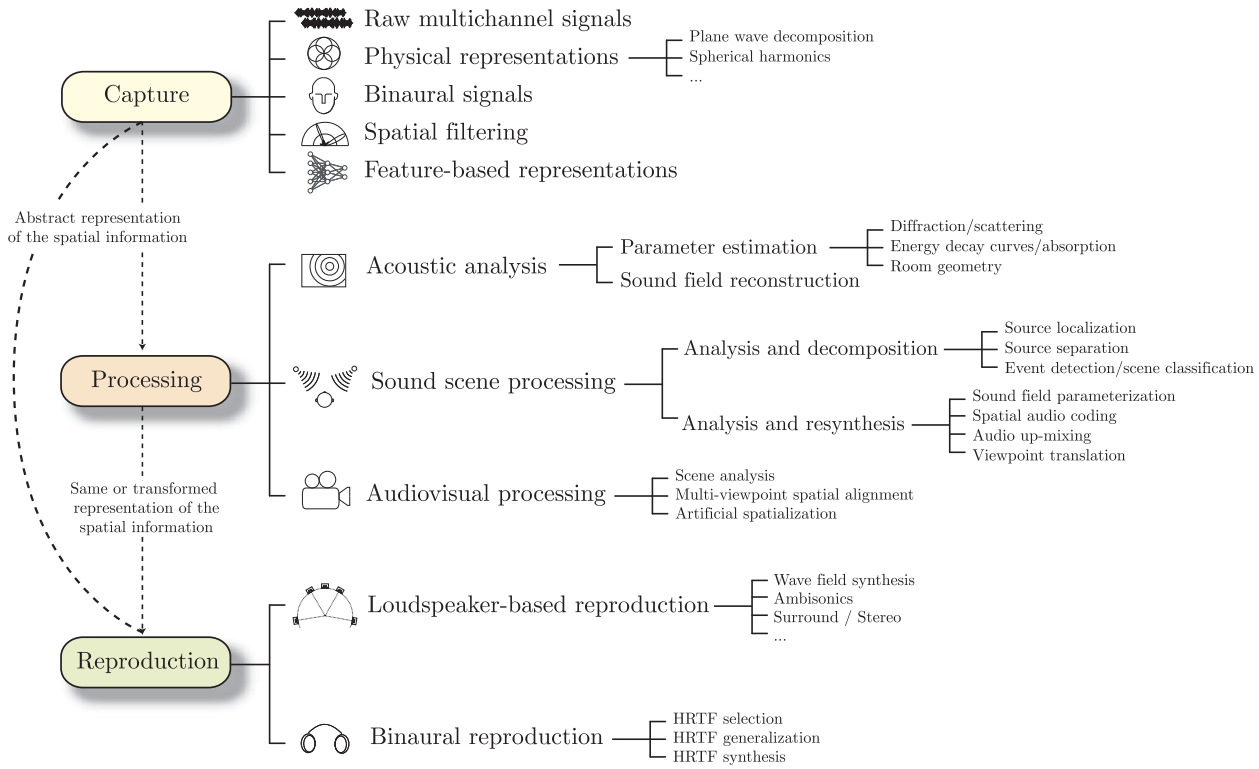

Fig. 1 A general conceptualization of the spatial audio pipeline

an abstract representation of the physical structure of the captured sound field. Examples for this are the plane wave decomposition or the spherical harmonic decomposition. The coefficients of the decomposition serve as the output format of the capture stage. The output of the capture stage may be piped directly to the reproduction stage, or, it may be processed.

The processing stage uses a suitable representation of the sound scene as input to, for example, extract information on the sound scene such as the number of sound sources and their locations or the instantaneous directions of incidence of the wave fronts. The processing stage may also manipulate the sound scene, for example, by separating direct sound components from diffuse reverberation and recombining them such that this results in a change of the characteristics of the reverberation or in a change of the apparent location of a source. The ultimate goal could be to decompose a sound scene into all its independent conceptual components, i.e., the individual source signals and all components of the reverberation that each source produces. This would allow for unrestricted manipulation. This goal still lies in a considerably distant future so that the available methods rather target different subsets of the sound scene components.

The reproduction stage renders the sound scene and produces the input signals to the loudspeakers that are available. These loudspeakers can either be mounted in a pair of headphones—one speaks of head-related reproduction—or mounted in the space around the listener(s)—one speaks of room-related reproduction [17]. Figure 3 presents some examples. A plethora of methods have been proposed for room-related reproduction depending on the number of loudspeakers that are available, the size of the listening area, and the number of simultaneous listeners [18]. Head-related reproduction injects the signals directly into the listener's ear canals and uses an acoustical model of the human head to convey the spatial information [19]. This acoustical model is represented by the user's HRTFs. As HRTFs are individual to a person, HRTF individualization also by means of ML has become a topic of considerable activity and is covered by this article.

Ideally, one would like to have available a universal representation of the sound scene based on which all conceivable methods of the processing stage can operate and that can serve as the input to the reproduction stage. As of now, such universal representation does not exist. Rather, a set of representations have become popular that are partly compatible and partly incompatible with each other so that many times; the employment of a given method in the processing stage poses certain requirements on the capture and/or the reproduction stage. Some

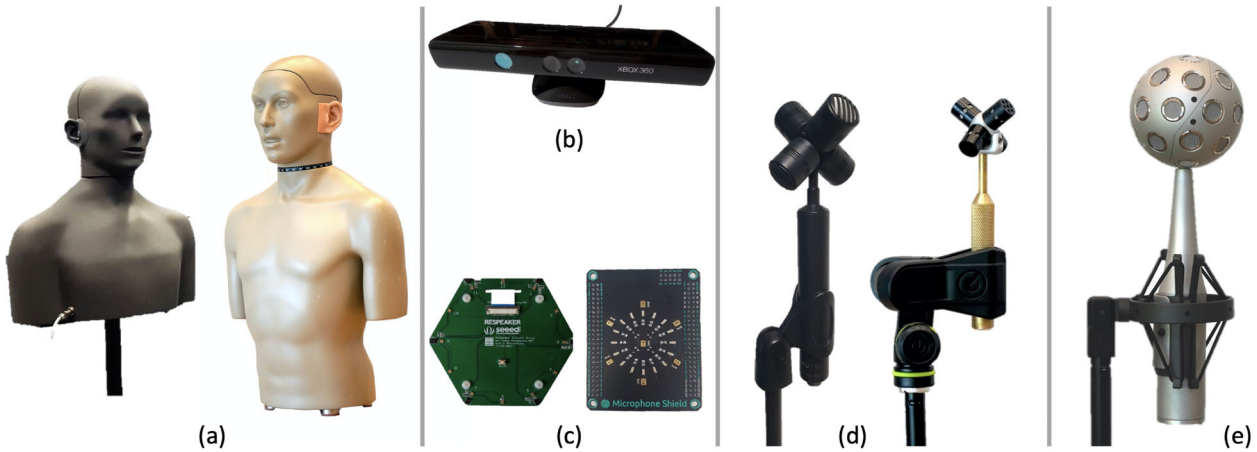

Fig. 2 a Dummy head recording systems Cortex Instruments MK1 (left) and GRAS Kemar (right), for binaural recording and psychoacoustical studies. **b** A consumer human-computer interaction device (Microsoft Kinect) equipped with a 4-channel linear microphone array. **c** Example on-board circular arrays employing MEMS microphones, the 6-channel ReSpeaker Circular Array (left) and the 7-channel MOJO Microphone shield (right). **d** Example tetrahedral arrays, the Rode Soundfield SPS200 (left) and Coresound Tetramic (right) for capturing first-order Ambisonics. **e** A high-resolution spherical microphone array, the 32-channel mh Acoustics Eigenmike

methods were formulated for processing exclusively spatial room impulse responses (SRIRs), i.e., room impulse responses (RIRs) that retain spatial information such as an array room impulse response, whereas other methods are formulated for running signals. We will not explicitly differentiate those as the underlying concepts are identical.

1.2 Data-based and model-based methods

As already introduced, the use of the term "data" can relatively easily lead to confusion within a conceptual framework merging traditional spatial audio concepts with ML. Note that ML-based algorithms are usually said to be data-driven approaches to emphasize the fact that they are designed to perform a given task by learning from data. In contrast, the term "data" in data-based spatial audio has been traditionally used to describe approaches that process signals in which the spatial information is encoded in the audio signals, even if the algorithms are not necessarily "data-driven." We illustrate the difference between data-based and model-based methods in spatial audio with a few examples in the following.

Stereophony uses differences between the two channels of a loudspeaker pair to encode spatial information [20, 21]. Typically, this is done using level or timing differences and also differences in the amount of signal correlation. Traditionally, the interchannel differences were produced by capturing a given scene with two microphones that are either located coincident and exhibit a suitable directivity or that are spaced to exploit differences in the arrival times of a given wave front. This may be considered a data-based representation as the spatial information of the sound scene is captured.

It is equally possible to create such interchannel differences manually by means of analog or digital signal processing so that spatial information can be imposed onto an otherwise non-spatial single-channel signal [21]. This may then be considered a model-based method.

Fig. 3 a Head-tracked dynamic binaural rendering with a head-mounted display (Oculus Quest). **b** Linear array of loudspeakers at Chalmers University of Technology. **c** Circular loudspeaker array suitable for 2D spatial audio rendering at University of Valencia. **d** Spherical loudspeaker array suitable for 3D spatial audio rendering at Aalto University

The underlying physical source model is not very explicit in stereophony. More recent spatial audio presentation methods like WFS or the ambisonics family can employ explicit physical source field models like spherical and plane waves [22]. Acoustic environments can, of course, also be represented using models [23].

Examples for modern purely data-based methods are the rendering of signals obtained from spherical microphone arrays (SMAs), which can be binaural [24] or using a method from the ambisonics family [22, 25]. Rendering of SMA signals has also been achieved in WFS [26].

Data-based and model-based rendering can, of course, also be combined, for example, by augmenting a data-based scene representation with additional model-based objects. Stereophonic recordings of large orchestras are simple examples where the spatial information from the main microphone is augmented with the signals from support microphones that are distributed across the orchestra and whose signals are panned manually to the desired spatial location. A more modern example are virtual panning spots that constitute virtual stereo loudspeakers that are embedded in a model-based scene and that contribute data-based information [27].

The named data-based methods all aim at maintaining the spatial information the way it was captured. The present article focuses on methods that go one step further in that they employ an enhancement of the original spatial information such as sharpening, up-mixing, or manipulation of the spatial information. This is performed in some cases in tangible representations of the data such as a plane wave expansion. In other cases, the data representations are abstract.

1.3 Article scope

Spatial audio covers a broad range of techniques and applications that encompasses a very large area or research. This overview intends to provide the reader with a comprehensive compilation of representative data-based methods having specific applications in spatial audio. While this is by no means an exhaustive description of all the existing approaches, the intention is to picture the current state of this field, elaborating on the impact that ML algorithms are having on different aspects of spatial audio research in the DL era.

Section 2 of the article focuses at the first stage of the pipeline, the spatial audio capture. The two major sound field representations (as opposed to array-specific representations) that are most commonly encountered in research, plane wave and spherical harmonic basis decompositions, are discussed in Sections 2.1 and 2.2, respectively. Some additional less common transform-based representations for the whole scene are discussed in Section 2.3, while Section 2.4 introduces the emergent topic of abstract acoustic and audio representations learnt directly from sound data with DL methods trained with a suitable task objective.

An overview of data-based spatial audio processing is provided in Section 3. The research topics involved herein span a very wide range of concepts, approaches, and applications, and the more mature methods that deliver spatial audio for reproduction are covered more extensively in Sections 3.2.2–3.2.5. Otherwise, advances in DL and other data-based methods that have strong potential in spatial audio applications are mentioned on the topics of acoustical analysis and parameter estimation in Section 3.1, signal decompositions or semantic descriptions of the spatial scene in Section 3.2.1, and joint audiovisual processing in Section 3.3.

In the subsequent Section 4, the final stage of the pipeline is briefly discussed. Tools for reproduction on loudspeakers or headphones of transform-based scene representations, spatial objects, or decomposed spatial components are technologically mature and are covered mainly by linear rendering techniques. Those are mentioned only briefly in Section 4.1 since they have been covered extensively in past literature. Recent DL-driven developments on personalization of spatial audio for headphone playback are reviewed more extensively in Section 4.2. Finally, in Section 5, we discuss the associations between the two major paradigms under review, signal-processing techniques, and the emerging DL methods, with regard to their potential in spatial audio, and we identify some open questions and possibilities for the future.

2 Spatial audio capture

Most methods for spatial audio processing, be it ML-based or not, use a scene representation as input that is either composed of "raw" microphone signals or originates from conventional linear processing applied to the signals from a microphone array. The goal of the capture stage is providing a representation that facilitates the application of a perceptual or physical model in the subsequent spatial audio processing stage.

There exist a number of general representations of SRIRs and measurement procedures for obtaining them that were proposed independent of a spatial audio context. General requirements for representations of room responses are identified in [28]. Measurement and extrapolation procedures based on sparse representations were proposed in [29–34] and based on non-sparse representations in [35]. The use of above methods in spatial audio is conceivable but has not happened on a large scale yet.

The main difference between these methods and those used in spatial audio applications is that the latter usually require information on the local propagation direction of a given sound field, which the referenced works do not comprise in a explicit manner. The remainder of the

section presents capture methods and the resulting scene representations that have been used explicitly in the field of spatial audio.

2.1 Plane wave decomposition

The plane wave decomposition (PWD) is primarily a fundamental mathematical representation of a general wave field. The circumstance that the basis functions are intuitive and even constitute useful conceptual elementary components of a sound scene made them popular [36]. An inconvenience in practice is that the PWD comprises parameters that are continuous with respect to space so that only a sampled (i.e., discretized) version can be stored and transmitted.

An early example for its use in spatial audio is [37], which converts a SRIR that was measured with a circular microphone array into a two-dimensional (2D) PWD for being able to auralise the space using WFS. The authors exploit the fact that it is known that the data represent a room impulse response by interpreting strong plane wave components as room reflections and rendering them in a dedicated manner per frequency band.

The microphone arrays employed in [37] are relative large in size with a radius in the order of 1 m. In [38], a similar method is proposed that employs compact microphone arrays that have a lower physical accuracy. The authors compensate for this limitation by using a higher degree of parameterization of the room response.

In [39], a method is proposed for manual manipulation of SRIRs that are parameterized in a manner similar to above described methods.

A variant of the classical PWD is the spatial decomposition method (SDM), which parameterises an SRIR into a single-channel pressure signal that encodes all temporal and spectral information as well as a direction-of-arrival (DOA) for each individual digital sample of the pressure signal. SDM was originally proposed for visualization of spatial room impulse responses [40]. An auralization of SDM-data both for binaural and for loudspeaker-based playback was proposed in [41]. Improvements of the loudspeaker-based variant were presented in [42] and in [43, 44] of the binaural variant.

2.2 Spherical harmonics-based representations

Another popular sound field representation is the spherical harmonic (SH) decomposition [45]. SHs are the angular solutions to the wave equation in spherical coordinates and are used in many fields of mathematics and physical science. Acoustic fields can theoretically be perfectly represented by superposition of an infinite set of SHs. In practice, a finite set has to be used, which limits the accuracy of the representation in different respects that are often abstract and intangible. Contrary to plane waves, SHs are a discrete representation, which means that a finite set of audio channels represents continuous spatial information. This aspect has contributed significantly to their popularity.

SHs found their way into the field of spatial audio through the visionary works of Michael Gerzon [46] and are often referred to as the ambisonics representation of a sound field. Later works particularly on spherical microphone arrays, much of which was performed by researchers from outside of the ambisonics community, highlighted the convenient properties of SH representations without necessarily referring to the concept of ambisonics [47, 48]. Even nowadays, the terminology is inconsistent in that many researchers do not necessarily employ the term ambisonics when dealing with SH representations of sound fields in a spatial audio context.

Particularly, the methods that will be discussed in Section 3.2.2 often employ an SH representation. In fact, while both PWD and SH representations may also be understood as acoustic analysis methods by themselves, this overview treats them as initial features that enable other spatial audio processing tasks.

2.3 Other transformation-based representations

Other representations based on linear spatial filtering techniques have also been proposed for a variety of spatial audio applications using space-time processing. Examples of particular interest are those that do not rely strongly on far field assumptions, but which approximate fields produced by nearby sources with far-field components. In this context, the ray space transform (RST) was proposed in [49] as a framework to formalize the plenacoustic analysis of [50, 51], through the adoption of Gabor frames. For its computation, the RST considers the output of a uniform linear array of microphones and applies a spatial sliding window to perform a "local" PWD of the recorded sound field. As a result, the RST is able to map the directional components of the sound field onto a "ray space" that provides some benefits in terms of invertibility and parameterization. For example, point-like sources are mapped onto linear patterns in the RST domain and spatial audio processing tasks such as source localization [52] or separation [53] can be directly performed over such representation. The projective form of the RST allows as well to process the signals captured by a set of compact microphone arrays, allowing applications such as sound field reconstruction [54].

2.4 Feature-based representations

The audio representations described above are all derived from mathematical manipulations that are both data-independent and motivated by already well-understood physical processes. Such representations have the advantage of being "general" and applicable to a wide range of problems, providing as well some valuable intuition on

the underlying acoustic phenomena. However, one of the most celebrated advantages of DL-based approaches is their capability to learn hierarchical representations of the input data automatically during training. Feature learning or representation learning is understood as a set of techniques that allow DL algorithms to discover automatically good representations from the input data that are able to encode efficiently the information needed for performing a given task. The feature learning process may be supervised (when labeled data is used) or unsupervised (when no labeled data is needed). In DL-based approaches, convolutional neural networks (CNNs) are typically applied to extract such abstract representations, which in the case of spatial audio should jointly capture spatial and spectro-temporal information about the sound scene. Since the majority of spatial audio processing techniques operate in the time-frequency domain, the most straightforward approach is to feed the network with the magnitude and phase of the available audio signals at the desired time-frequency resolution, and let CNNs extract the relevant information needed for accomplishing the task, as identified by their internal activations. Classical representations such as SHs can also conveniently be used as input features.

Another approach is to provide "hand-crafted" features that already represent meaningful information for the intended task. For instance, spatial information can be conveniently represented by sines and cosines of inter-channel phase differences [55, 56] or generalized cross-correlations between the audio signals [57–59], which helps to avoid phase wrapping problems and thereby eases the network training. Sound intensity computed from the SH representation, which has been successfully applied in DL-based sound localization in 3D [60], is another example of such a "hand-crafted" feature obtained in a pre-processing step. Another recent example is that of [61], where a rotation-invariant DNN architecture that performs sound event localization and detection was proposed for SH signals.

Spatial audio methods and systems rely on well-known perceptual mechanisms used by the auditory system [19]. Spatial hearing cues result from the acoustical interaction of an impinging sound and a listener's anthropometric features, which leads to filtering effects caused by the head, shoulders, torso, and pinnae. Interaural differences also have a strong influence. The above cues are typically encoded into HRTFs in the frequency domain or, equivalently, into head-related impulse responses (HRIRs) in the time domain. Datasets of HRIRs measured over a grid of spatial locations are important not only for realistic reproduction purposes but also for extracting general or universal patterns useful for understanding the relative influence of certain spectral features in the perceived sound. Traditionally, studies aimed at analyzing spatial audio perception have relied on listening experiments, which usually require a carefully designed and time-consuming experimental setup.

The data extracted from HRIR measurements has been computationally analyzed in the past by ML algorithms to gain insight into the auditory localization process. One of the earlier attempts was [62, 63], where a biologically inspired model of the source localization process was conceived by combining a cochlear model and a time-delay neural network. Similar ideas have been more recently exploited with the advent of DNNs. In [64], a spiking neural network is used to extract features from binaural recordings and training a three-hidden-layer feedforward network on such features to perform both single-source and multi-source localization over a range of noise conditions.

The learning capabilities of CNNs were recently exploited in [65] to identify primary elevation cues encoded in HRTFs shared across a population of subjects. A CNN was trained on multiple HRTF datasets to estimate the elevation angle of a virtual sound source, and salient audio features were extracted by using layer-wise relevance propagation. The results indicated that the discovered features were in line with those obtained from the psychoacoustic literature.

The spatial information comprised by binaural signals has also been exploited by DNN-based approaches to understand the spatial arrangement of musical acoustic scenes. In [66], a CNN was trained to classify binaural music recordings into foreground-background, background-foreground, and foreground-foreground scenes, indicating the relative position of the listener with respect to ensembles of musical sources (foreground) and room reflections (background). The authors compared the performance of automatic algorithms to that of human listeners in this task [67], with results suggesting that ML algorithms can significantly outperform human listeners under matching binaural room impulse reponse (BRIR) conditions (test scenes rendered by using the same set of BRIRs as training scenes) and exhibiting similar performance in the mismatched case. Despite the task addressed is not particularly aimed at feature discovery, meaningful internal representations might be obtained as a byproduct.

Techniques such as the ones described above and others that also exploit visual information (cf. Section 3.3) may open the door to learning-based methods capable of leading to alternative signal representations for spatial audio.

3 Spatial audio processing

The processing stage typically either extracts desired information from the output of the capture stage or manipulates the spatial information. Indeed, most meth-

ods for spatial audio processing are data-based in nature, as their ultimate goal is usually aimed at analyzing or modifying the spatial information present in their input signals or, alternatively, to the exploitation of such spatial information to extract meaningful constituent signals of the sound scene. Signal enhancement by modifying the statistical relation between signal channels may also be considered part of the processing [68]. Some of the methods presented in this section can be used sequentially in a processing pipeline. Many reference and commercial methods for spatial audio processing today are still based on classical signal processing, as in the case of the family of parametric spatial audio techniques (Section 3.2.2) or the methods employed within spatial audio coding standardization frameworks (Section 3.2.3). This is also the case for the vast majority of viewpoint translation methods (Section 3.2.5). On the other hand, many DL-driven approaches have recently appeared in the context of acoustic analysis (Section 3.1.1), sound scene decomposition (Section 3.2.1) and audio-oriented audiovisual processing (Section 3.3), which are rapidly displacing traditional methods. For example, DL approaches are now a reference in fields like source separation and enhancement. In a middle-point, while classical methods are a reference for audio up-mixing due to their inherent relation to parametric spatial audio approaches (Section 3.2.4), there is a clear trend in the use of DL for such task. As a result, we will cover all the above spatial audio processing systems by emphasizing such diversity and coexistence of traditional and ML-based techniques.

Spatial audio processing techniques can be of a very different nature and oriented towards significantly different objectives. For the sake of clarity in the presentation, we broadly divide processing techniques into three blocks (cf. Fig. 1). The first one covers techniques aimed at analyzing and describing acoustically the sound field (Section 3.1). The second block describes techniques for sound scene processing with a special emphasis on methods oriented towards the analysis and modification of the spatial information in the recorded scene for subsequent re-synthesis (Section 3.2). Finally, we discuss recent approaches making use of audiovisual data to address several tasks related to spatial audio in the third block (Section 3.3).

3.1 Acoustic analysis

This section covers data-based spatial audio methods for acoustic analysis. We discuss separately methods aimed at estimating acoustic parameters for sound field rendering and those for acoustic imaging and sound field reconstruction. While a large body of literature exists on these topics, we limit our discussion to the most recent approaches that make use of DL. Note, however, that in order to provide a comprehensive overview of such recent approaches, we cover as well works where visual data is considered as input, even though the final objective is on the acoustics side.

3.1.1 *DL-driven acoustical parameter estimation for spatial audio rendering*

The capability of DL to model complex relationships between different representations and their effects in a certain domain has found recent use in acoustical modeling problems. An interesting such application is on acoustical parameter estimation for virtual acoustics and spatial auralization. More specifically, a DNN is trained in [69] to map a rectangular plate geometry that occludes a source from a receiver to filter parameters modeling the perceived effect of the diffracted sound at the listener. Going even further, fast computation of the 2D scattered field around an acoustically hard 2D object is approached in [70] as an image-to-image learning task for a CNN, trained on images generated with wave-based acoustical simulations. The principle is inverted in a further work to estimate the 2D shape of the object from its scattered field in [71]. Finally, a similar training strategy is followed in [72] while instead of 2D objects and field images, they map 3D geometries to far-field spherical harmonic coefficients of the scattered field. Note that training NNs to model acoustical scattering [73] or infer geometry from scattering measurements [74] has been attempted much earlier than those works. However, their considerations are different.

Acoustical parameter estimation using DL has also been used to extract room acoustic parameters from geometry or image data for fast spatial rendering in audio VR. In [75], energy decay relief curves are estimated directly from images of acoustical spaces using CNNs. Training pairs of features of room geometrical configurations and spatial impulse responses captured for those configurations are used in [76]. Alternatively, simplified room geometries are reconstructed from 360 camera images in [77, 78], which are used to drive virtual acoustic simulators for AR/VR applications. In the same spirit, [79] uses a DNN to classify materials from textures in a 3D room geometry, to deduce and optimize absorption coefficients to be used in conjunction with the geometry for interactive geometrical acoustics simulations. Finally, [80] combines measurements with geometric generation of reverberation by estimating a simplified geometry from a moving 360 camera recording with structure-from-motion and used to synthesize early reflections at any position. Additionally, a single monophonic RIR is captured in the room and used as a guide for obtaining absorption filters for the inferred geometry, low-frequency modal filters, and also the late reverberation tail to append to the generated early part in a position-independent manner. The method is used to synthesize ambisonic SRIRs. The work is extended

in [81] by replacing the RIR measurement signal with a general audio signal such as speech and using commodity hardware (i.e. a mobile phone). Since the RIR parameters are not readily available in this case, they are estimated from the source recording using DNNs.

3.1.2 DL-driven sound field reconstruction

Recently, DL techniques have also been exploited for sound field reconstruction from a small number of irregularly distributed microphones in a room. The work in [82] proposed the use of a U-net neural network [83] with partial convolutions trained on simulated data for sound field reconstruction in rectangular rooms. The proposed data-driven method allows to reconstruct the magnitude of the sound pressure on a plane, performing jointly inpainting and superresolution from irregular discrete measurements in the frequency range 30–300 Hz. The same method was recently extended to reconstruct both magnitude and phase with testing over real-world sound fields in [84] using a publicly available dataset, showing as well the potential application of DL-driven sound field reconstruction in sound zone control. Other interesting and novel architectures are appearing in the context of sound field analysis for acoustic imaging. In [85], a recurrent neural network with a fully customized architecture is proposed, taking into account relevant aspects of acoustic imaging problems and inputs from a spherical microphone array. In this context, other DL-driven approaches for high-accuracy acoustic camera solutions have recently appeared, which also make use of additional modalities such as stereo vision technology [86].

3.2 Sound scene processing

We refer to sound scene processing techniques as those that analyze and manipulate audio signals with the aim of extracting and modifying the spatial information in the captured scene by decomposing the scene into perceptually meaningful elements. These elements may be either in the form of spatially relevant components (e.g., directional vs. diffuse sound) or related to the sources making up the scene. Despite its proximity to the term "auditory scene analysis" [87, 88], coined by psychologist Albert Bregman and usually linked to the field of source separation, we use the term sound scene processing in a more general way that encompasses not only the extraction and decomposition of sound into source objects, but also the analysis and manipulation of spatial features.

3.2.1 Sound scene decomposition

With the term scene decomposition, we refer to the wide variety of methods that aim to break down the sound scene into its constituent components. A prominent such example is decomposing the scene into constituent signals based on their spatial properties, such as foreground-background, primary-ambience, or directional and non-directional separation. More elaborate spatial decompositions can detect and separate distinct localized sources from different directions. Such decompositions rely heavily on source detection, source localization, and spatial filtering techniques. These research topics constitute core problems in microphone array processing with an accumulated intensive research history of decades and applications spanning a much wider range than the scope of this article. Of course, such techniques are employed by, e.g., the parametric spatial audio processing methods reviewed in the next section, but they are not analyzed separately. For a comprehensive overview of them the reader is referred to [89, 90]. Recently, there is also intensive research on DL variants of source localization, e.g., [91–93], and spatial filtering [94].

Another family of methods aiming to decompose the spatial scene into its constituent signals is termed multichannel source separation. Many examples are closely related to adaptive and informed spatial filtering, as reviewed in [90]. However, while localization and spatial filtering are used extensively in spatial audio methods, the source separation research has generally focused on maximum separation of source signals, and not on re-rendering or modifying spatially the scene. However, a stronger separation component has obvious applications in spatial audio, such as spatial remixing of the scene and other source-dependent modifications. Works that follow a source separation perspective for spatial audio can fall into two categories. The first aims to recover a demixing matrix or separation masks using mainly spatial features and a mixing model in the time-frequency domain, such as the works in [95–97]. The second category attempts an even higher-level decomposition of the scene, integrating apart from spatial mixture models, also spectrotemporal models that distinguish one source from another. Separation in this case can be performed blindly or in a supervised manner, using some prior information on the spectral templates of the sources in the scene. Only a few works have attempted applications of those models to spatial audio rendering [98, 99]. In general, multichannel source separation is transitioning very quickly to DL-driven solutions, which, however, are currently focused on multi-speaker separation and enhancement rather than scene audio modification, or resynthesis [55, 100, 101].

Apart from signal decompositions, higher level audio scene analysis with semantic information is an extremely active field of research that is completely dominated by data-based DL approaches [102]. Examples include acoustic scene classification (ASC) [103, 104] and simultaneous temporal detection and sound-type classification of multiple concurrent sound events in the scene (commonly known as sound event detection (SED) [105, 106]). A large part of this research community participates in the DCASE Workshop and the associated DCASE Challenge

[103, 105, 107]. Interestingly, this research community had not involved spatial information until recently, with a few exceptions such as [104] in ASC or [106] in SED. However, currently there is increased interest on advanced spatiotemporal semantic descriptions of the sound scene, with the common task of jointly performing sound event localization and detection (SELD) using multichannel signals [91, 107]. Semantic descriptions and decompositions of sound scenes are of course of interest also in spatial audio analysis and synthesis, and stronger cross-pollination between those research directions and spatial audio is expected to take place in the coming years.

3.2.2 Sound field parameterization for analysis, modification, and resynthesis

In this section, we present an overview of parametric spatial audio techniques that analyze the captured sound scene to obtain a compact yet perceptually relevant parametric representation and subsequently re-synthesize it for spatial reproduction. In these techniques, the estimated signals together with supplementary parametric information serve as a basis for modifications (processing) as desired by the target application. A description of many state-of-the-art approaches to parametric spatial audio processing can be found in [108].

The stepping stone in development of these methods was the proposal of spatial impulse response rendering (SIRR) [109, 110], which processes SRIRs in ambisonics B-format (i.e., a 1st-order SH representation) to decompose them into one direct-sound component and a diffuse residual for each time-frequency bin. The underlying assumption is that source signals tend to be sparse in the time-frequency domain (for example, the energy of a periodic signal concentrates at the harmonic oscillations) so that a single wave front sufficiently represents the direct-sound component at a given frequency bin [111]. This concept was extended in [112] to running signals and was termed directional audio coding (DirAC). In DirAC, a zeroth-order (omnidirectional) signal is supplemented by two parameters, namely the diffuseness and the DOA of the direct signal. The former is estimated from the temporal variation of the intensity vector [112, 113], and it is used to extract the direct and diffuse signal components from a mono signal using a single-channel filter [112]. In later work [114], the parametric approach was extended to arrays of any type whereby the extraction of the direct and diffuse signals is typically performed using signal-dependent spatial filters, many of which are well-known from speech enhancement [89, 90, 115]. The extracted signals are supplemented by parametric information on the DOAs or the positions from which the direct sound components originate.

Examples of parametric modifications include rotations of the entire recorded sound scene or manipulations of the locations of individual directional sounds [25, 116, 117]. For reproduction of musical recordings, the diffuse signal is usually subject to decorrelation before it is fed to the loudspeakers in order to increase the feeling of spaciousness and plausible listener envelopment [112, 118]. Another example is increasing quality by a reduction of coloration while providing stable localization cues when using recordings of spaced microphone arrays [119]. Furthermore, when capturing the acoustic scene using distributed microphone arrays, the signals to be reproduced can be synthesized for an arbitrary listening position in space. This can be achieved by synthesizing the signals of virtual microphones of arbitrary spatial patterns at locations that are not populated with real microphones [120, 121]. Similarly, binaural signals for different virtual listening positions can be synthesized [114], which we discuss in some detail in Section 3.2.5. In teleconference applications, preservation of spatial cues combined with flexible spatial selectivity offered by parametric approaches can help the auditory system to naturally focus on a desired speaker [114, 122], which may lead to better speech intelligibility. By adjusting the output parametric gains for the direct and diffuse signal components, it is also possible to align the visual and acoustical images in digital camera recordings, including the effects of an acoustical zoom that is consistent with the visual cues and a blurred spatial audio image for sources located off the focal plane [123, 124]. Another approach to generate binaural or multichannel audio which follows the moving picture of a visual scene is to perform adaptive equalization of the direct signals [125].

Considerable research has also been carried out to extend and improve the parametric representations. Early attempts include the higher angular resolution plane wave decomposition (HARPEX) [126], which decomposes B-format signals into two plane waves per time-frequency bin, and a method in [127] in which the higher-order signals are decomposed into several plane waves, while the diffuse residual is in both cases ignored. Higher SH or microphone orders (HO) enable higher spatial resolution, which allows differentiation of more than one simultaneously impinging wave front. SIRR was extended to HO-SIRR in [128, 129]. DirAC was extended to HO-DirAC in [130] whereby the standard parameterization is performed separately for a set of angular sectors to support several directional and diffuse sounds arriving from spatially separated directions simultaneously. More recently, coding and multidirectional parameterization of ambisonic sound scenes (COMPASS) [131] extends the parametric model to several overlapping directional sounds and a diffuse residual per time-frequency bin, and it also provides a convenient method to combine the parametric processing with standard ambisonic reproduction.

3.2.3 Spatial aaudio coding

The delivery of spatial audio content to the masses and the transfer of academic research to the consumer and media industries involves making such delivery flexible and efficient. The adoption of spatial audio formats and processing schemes within standardization activities should not be ignored in this overview, as they can have a very significant impact on how spatial audio will be consumed at scale and the momentum that spatial audio technology may experience in the coming years. The development of audio applications for consumer electronics and multimedia streaming services brought about a demand for representing spatial audio content in formats that support efficient transmission at limited bandwidth and require scarce storage capacity. Over the last 20 years, a number of spatial audio coding techniques have emerged for 2D and 3D reproduction that to a large degree maintain the fidelity of the rendered spatial scene. Although ML has not yet been incorporated into popular spatial audio coding formats, we can already observe an interest in applying ML as part of the processing chain, especially for data compression at low bitrates.

Early coding standards include Moving Picture Expert Group (MPEG) parametric stereo [132, 133] and Dolby Prologic, in which spatial information is encoded by manipulating the phase differences between the stereo channels. A notable successor has been MPEG Surround [134], which exploits three major spatial cues attributed to the perception of the 2D spatial sound image, namely inter-aural level differences, inter-aural time differences, and inter-aural coherence [135–137]. These perceptual spatial cues have formerly been applied in binaural cue coding [136, 137]. The encoder of MPEG Surround extracts spatial information such as channel level differences and inter-channel correlations from pairs of input audio channels using an complex-exponential modulated quadrature mirror filter (QMF) filterbank. Together with additionally estimated channel prediction coefficients and prediction residual signals, this side information is transmitted along with the down-mixed mono or stereo signals to a decoder that uses it for up-mixing to multichannel audio. A stepping stone in developing high-fidelity spatial audio codecs has been the MPEG-H 3D Audio [138, 139] standard, which supports spatial coding of multichannel audio signals, sound objects, and HO ambisonic signals. Similarly to the parametric methods discussed in Section 3.2.2, the scene can be decomposed into sound objects that are either static or their positions and gains may vary over time, as in MPEG-D spatial audio object coding (SAOC) [140]. The remaining ambience sound field components can be conveniently coded in an HO ambisonic representation. Until recently, attempts to incorporate ML into spatial audio coding have concentrated predominantly on the inference and compression of the associated parametric side information. In [141], ML is employed in visual and audio-visual tracking of sound objects in an end-to-end audio-video approach. More recently, DL has been applied to spatial audio object coding [142], in which a mixture network of deep convolutional architectures enable to effectively compress spatial parameters of audio objects at low bitrates.

Note that apart from coding of spatial information, a significant part of the discussed codecs concerns audio signal compression. For instance, compression in MPEG-H 3D audio is performed based on MPEG-D unified speech and audio coding (USAC) [143], while MPEG-4 high efficiency advanced audio coding (HE-AAC) [144] is typically used in MPEG Surround. The virtue of using ML in audio compression has been demonstrated in [145] for extending the frequency bandwidth and in [146] for reducing lossy coding artifacts. However, a full end-to-end neural audio codec has only very recently been proposed in [147]. The DL-based SoundStream model, which is composed of a fully convolutional encoder-decoder structure and a residual vector quantizer, has been designed to provide good quality at extremely low bitrates. Extensions of ML-based audio compression to ML-based spatial audio coding are well expected, yet they are still to come.

3.2.4 Up-mixing

Up-mixing techniques are those aimed at generating a higher number of audio signals from a smaller set of audio channels while preserving important aspects such as the locations of the main sound sources or the ambience components contained in an original sound recording. This comes usually with an apparent increase of the spatial resolution. As a result of the up-mixing process, recordings coming from a down-mixing process can be automatically extended to multi-channel arrangements with the objective of conveying a more natural and enveloping experience.

The sound field parameterization methods presented in Section 3.2.2 exhibit such functionality inherently, for instance, by reintroducing the extracted direct sound into the scene representation with a higher SH order such as in [42, 43]. Based on the compact parametric representation, these methods can synthesize the loudspeaker channels of arbitrarily high orders by means of panning the directional signals and decorrelating the diffuse residual signals to be played back over all loudspeakers. However, the most popular application of up-mixing is in spatial audio coding covered in Section 3.2.3, where at the decoding stage, multichannel loudspeaker signals are synthesized from down-mixed representations. Several data-based methods have been proposed for up-mixing from mono or two-channel stereo to multichannel [148]. For up-mixing to a 5.1 format in MPEG Surround, the powers and cross-correlations of stereo signals are analyzed in perceptually

motivated sub-bands to extract coherent signals to be played back using a pair of front loudspeakers that enclose the estimated direction (i.e., left and center or right and center) as well as to extract the lateral ambience signals to be emitted from side or even all loudspeakers for improved listener envelopment. In [149], ambience components are identified and extracted based on inter-channel coherence first. Then, they undergo decorrelation using all-pass filters in order to avoid undesired phantom images to the sides of the listener. The panning gains of individual directional signals are found based on the so-called inter-channel similarity measure, and a nonlinear mapping is applied to re-pan the sources from 2 to 3 front channels for improved stability of the spatial image for off-center sources. By introducing more than three front loudspeakers, the width of the sound scene can be further increased beyond the standard $\pm 30°$, and the listening sweet spot region can be increased. In [68], an adaptive mixing solution was proposed to reach the target covariance matrix by exploitation of the independent components in the input channels, while minimizing the usage of decorrelated ambient signals when the target covariance cannot be reached without the application of decorrelation.

Recently, also DNNs have found application in the development of audio up-mixing and surround decoding systems. One of the earlier attempts for ambient extraction from mono signals using neural networks was proposed in [150], where a shallow architecture with one hidden layer was used to estimate spectral weights relating the ratio between the ambience and direct signal components in the time-frequency domain. The input to the network were well-known hand-crafted audio features such as spectral centroid, spectral flatness, or spectral flux. More recently, a DNN-based method to process stereo tracks was proposed in [151] that is aimed at classifying and separating the primary (direct) and ambient (diffuse) components in each time-frequency bin of the input mixture. In this case, a feedforward network with three hidden layers and a sigmoid-activated output layer was used for classification, building a time-frequency mask for the subsequent separation. Another work exploiting DNNs for stereo to 5.1 up-mixing was proposed in [152] considering the MPEG-H 3D framework. In this approach, DNN models for the center and surround channels are trained by using log-spectral magnitudes of QMF subbands. The input stereo signals are converted into rear and center channels using the trained models, where the generated subband signals are transformed back to audio signals using QMF synthesis. Following a similar subband approach, the authors proposed in [153] a method for converting mono signals to stereo training multiple DNNs for each sub-band with the objective of modeling the band-wise nonlinearity between the mid and side signals.

The system proposed in [154] uses two networks for two-to-five channel up-mixing, where one of the networks is used for primary and ambient signal separation and the other for ambience rendering. The networks are jointly trained by minimizing the mean-squared error between the magnitude spectra of the original and the decoded five-channel signals as well as the interchannel level differences of the target signals. The obtained spectral weights are multiplied for each frequency bin of the input stereo signal, allowing for the separation of primary and ambience signals and the generation of diffuse sound, respectively.

3.2.5 Viewpoint translation

The advent of VR and AR goggles has boosted the interest in both academia and industry in binaural rendering technologies. The two most dominating fields of activity in this regard are HRTF personalization (cf. Section 4.2) and 6-degree-of-freedom (6-DoF) binaural rendering. This section focuses on the latter. The 6 DoF in this case are 3 angles of head orientation as well as head translations in the 3 Cartesian dimensions. 6-DoF binaural rendering obviously requires tracking of the user's orientation and position in realtime on the rendering side. The requirements for the performance of the head tracking that VR and AR goggles perform for the visual rendering, particularly signal-to-noise ratio and low latency, are much stricter than what is required for the audio rendering so that tracking is readily available.

Most methods employ a SH representation of the sound field that is to be rendered. Rotation of this representation relative to the HRTFs that are used for the rendering is straightforward. 6-DoF rendering is achieved in [155] via the application of blind source separation to the captured sound field. A method based on DirAC is presented in [156, 157]. Translatory head movements based on a plane wave expansion of the sound field to be rendered was presented in [158, 159], which demonstrated fundamental limitations of this framework. As a consequence, 6-DoF binaural rendering methods typically use a more application-oriented sound field representation and come in four flavors: (1) methods that perform a mathematical translation of the orthogonal sound field decomposition or a re-expansion around a different center [160–163], (2) parameterization and adaptation of a single-viewpoint recording (or RIR measurement) with a microphone array [116, 117, 164–166], (3) interpolation between microphone array recordings performed at different locations [167–170], and (4) interpolation between parameterizations of recordings performed at different locations [171–175].

3.3 Audio-visual processing

Some of the methods described in Section 3.1 make use

of visual data to gather supplementary environmental or geometric information, which assists the estimation of acoustic-related parameters. Combined audio and video analysis has been performed extensively by the computer vision community, providing stronger cues for, e.g., activity detection or speaker recognition than processing video or audio separately. Some of the studied tasks overlap with audio-only tasks such as on-video speaker and sound source localization [176] and video-guided monophonic source separation [177]. In any case, there is no doubt that the use of video recordings for addressing spatial audio tasks has been receiving increasing attention in the last years.

In this context, the availability of very large audio-visual datasets have contributed significantly to promoting the use of audiovisual information to exploit both auditory and visual spatial cues jointly. In [178], a dataset of 360° videos from YouTube containing first-order ambisonics audio was collected to train a self-supervised audiovisual model aimed at aligning spatially video and audio clips extracted from different viewing angles. The approach was shown to yield better representations, outperforming prior work on audio-visual self-supervision for downstream tasks like audio-visual correspondence, action recognition, and semantic video segmentation. Similarly, the work in [179] proposed another self-supervised approach to understand audio-visual spatial correlation by training a DNN over a large dataset of ASMR (autonomous sensory meridian response) videos to classify whether a video's left-right audio channels had been flipped. The learnt audio-visual representations were proven to be useful for carrying out some downstream tasks including source localization, mono-to-binaural upmixing, and sound source separation. A similar application is addressed in [180], where a U-net network is proposed to infer binaural audio from videos and their respective monophonic audio recording using a database of binaural music recordings as training targets. Similar to mid-side stereo, a mid signal corresponds to the mono mix-down of the binaural audio, and the network learns to predict the side signal only. A similar approach was followed in [181]. In [178], a network is taught to upscale a monophonic signal to first-order ambisonics by self-supervised learning from 360° videos. The network is taught to produce time frequency masks to separate the mono input into directional components along with a set of directional weights encoding those components into first-order ambisonics signals. The work in [182] shares some similarities but assumes static source positions and does not perform source separation. Finally, an attempt of a completely synthetic approach to generation of an audible scene from a 360° image is presented in [183] by mixing background ambience based on scene classification and object, people, or action sounds, based on visual object recognition and spatialized at their respective detected image locations. Of course, there is no temporal information on the arrangement of events in this scenario, bringing the work closer to sonification of the immersive image.

4 Spatial audio reproduction

The final stage in the spatial audio pipeline is the reproduction of the multichannel signals that result from the preceding capture and processing stages. In general, the theory underlying spatial audio reproduction is well established and no disruptive methods or discoverings have been observed in the last years, especially in the context of loudspeaker-based reproduction. It is true that the power of DL has attracted great interest in the context of binaural reproduction where some interesting DNN-based approaches have recently emerged for selecting or synthesizing binaural signals adapted to a given listener. Therefore, while this section briefly outlines conventional linear spatial audio reproduction methods for the sake of completeness, only the aforementioned DL-driven attempts are reviewed.

4.1 Linear spatial audio reproduction

Any audio reproduction method converts a scene representation into loudspeaker input signals that produce a sound field with a given desired physical structure or binaural signals with given desired properties. Traditional audio reproduction concepts like stereophony feed the signals from the microphones of the capture stage directly into the loudspeakers of the reproduction stage. More advanced concepts like modern ambisonics formulations perform linear filtering operations to compute the loudspeakers signal from the entire set of microphone signals whereby one of the scene representations discussed in Section 2 can be an intermediate format. Many such methods are linear.

We refer the reader to the literature such as [184, 185] on binaural rendering, [122, Ch. 14] on loudspeaker panning, [25] on ambisonics, and [186] on wave field synthesis for more detailed discussions. Further overviews are provided in [2, 17, 18, 187]. We will focus on certain aspects of binaural reproduction the subsequent Section 4.2. Binaural rendering in its most common form is a linear method that is straightforward and comprises filtering the given scene representation with a suitable representation of the user's HRTFs. What is interest in the scope of the present article is the computation of HRTFs for this purpose that are personalized to the user by means of data-based processing.

4.2 DL-driven HRTF personalization and generalization

HRTFs are highly dependent on the individual anthropometric features of a given listener. A major challenge for

personalized binaural audio reproduction is the measurement procedure of HRTFs, which is tedious and expensive. Research efforts have therefore addressed the problem of HRTF customization with the objective of estimating individual HRTFs based only on geometric information or user feedback without the need for any measurement process. Traditionally, principal component analysis (PCA) has been the technique of choice for dimensionality reduction in HRTF datasets, leading to many interesting observations and experiments that evaluate the impact of different eigenmodes on spatial audio perception [188]. Further improvements in PCA-based HRTF modeling and customization were presented in [189, 190]. HRTFs were synthesized in [191] using a sparse combination of a subject's anthropometric features. The use of DNNs in the subject matter is rapidly gaining momentum.

An HRTF selection method based on a multi-layer perceptron neural network was proposed in [192]. The system was trained by using as input a set of measured anthropometric parameters (shapes and sizes of listeners' heads and pinnas) extracted from photographs. To train the network, 30 subjects listened to music rendered by using different HRTFs to assess their fitness and obtain a score used as target output. Such an approach was shown to be more effective in selecting the best matching HRTF for a given listener than selecting the one with the smallest sum of squared errors between the listeners' measurements and each of the database members. In a similar spirit, but with the aim of synthesizing a personalized HRTF, [193] proposed a method consisting of three sub-networks: a feedforward network taking as input anthropometric measurements, a CNN using ear images, and another feedforward network that estimates a personalized HRTF by using the outputs of the other two subnetworks.

Autoencoder (AE) architectures have been selected by several works to capture relevant patterns across HRTF datasets. An AE is an unsupervised DNN that learns how to efficiently compress and encode data by learning how to reconstruct the data back from a reduced encoded representation, usually referred to as embedding. In [194], a sparse AE is used to create embeddings from the captured HRTFs, which are used to train a generalized regression neural network (GRNN) to approximate equivalent latent representations of the corresponding HRTFs at arbitrary angles. The sparse AE is then able to decode the GRNN output to reconstruct the desired HRTF. Such an approach provides an efficient way to jointly address the creation of generalized HRTF models and angle interpolation from large datasets. Similarly, a set of independent AEs was used in [195] for each elevation angle to reconstruct HRTFs on the horizontal plane and used the resulting latent representations in the bottleneck as targets for a feedforward network using anthropometric features as input. A personalized HRTF can then be synthesized by estimating the latent representation given the features of a new subject and feeding the result into the decoder part.

In [196], a training and calibration procedure based on a variational AE structure was proposed. Variational AEs are deep generative models that provide a probabilistic manner for describing an observation in latent space, modeling the probability distribution of the input data. In the training step, an HRTF generator is created by learning the individual and nonindividual features from an HRTF dataset. The generator is based on an extended variational AE that separates with a set of adaptive layers the individuality and non-individuality factors of the users in a nonlinear space. The learned latent variables together with some personalization weights optimized by user feedback are then used in the calibration step to generate a personalized HRTF for a specified direction.

A study of several aspects related to HRTF individualization that provides further insight into the research lines discussed above is presented in [197]. As expected, models seem to generalize better by having access to larger datasets. This may be achieved by a proper and careful merging of existing datasets [198] or by synthetically creating new ones [199].

5 Discussion

Many of the paradigms behind spatial audio are deeply rooted in physics, whereby the ultimate goal of spatial audio is evoking a given sensation in the listener. This sensation may be evoked through pure physical accuracy, i.e., if one intends to reproduce a performance in a concert hall, then the perfect re-construction of the sound field in the concert hall at the listening location is guaranteed to provide the best possible perceptual result. But it has been clear since a long time that such physical accuracy would require immense resources, if it is achievable at all [22, 200]. It was shown in different situations that authenticity, i.e., a reproduced scene being indistinguishable from the captured original scene, can be achieved even if the physical accuracy is relatively low [201–203]. The interpretation of the ear signals by the human auditory system seems to lead to the same perceptual result for different input signals in certain situations. This psychoacoustic route is what virtually all methods in the processing stage have been taking. Some concepts that are applied in the capture and reproduction stages also rely explicitly or implicitly on given psychoacoustic properties of the human hearing system [187].

The capture and reproduction stages base heavily on the underlying physics of the problem, such as the relation between the signals impinging on a set of microphones or the spatial structure of a sound field created by an array of loudspeakers. These physical mechanisms are well understood, and there exist powerful mathematical tools that describe those in a compact and accurate way.

The non-ML data-based approaches in the processing stage use perceptually motivated representations of acoustic scenes to accurately reproduce the relevant spatial cues while preserving the highest audio signal fidelity. Many of these data-based techniques, described, e.g., in Section 3.2, including parametric processing, up-mixing, spatial audio coding, and viewpoint translations, successfully achieve the designated goals. As a result, these methods steadily play a dominant role in present audio and multimedia applications.

The recent irruption of DL has opened new opportunities for spatial audio, offering enticing alternatives to the classical signal processing. Major breakthroughs in DL have taken place in the context of image processing, and these advances have been quickly adopted in closely related speech and audio tasks. For instance, image segmentation relies on partitioning a digital image into meaningful segments such as objects or contours by assigning a corresponding label to every pixel in an image. Acoustic source separation can be considered an analogous task where source labels are assigned to each time frequency bin in a 2D time frequency representation of an audio channel. Similarly, image classification in vision can be considered an analogous task to acoustic scene classification in audio since both tasks assign a class label to the input signal representation, which is clearly evident when a 2D time frequency audio spectrogram is treated as an input image.

Due to the apparent similarities, DL approaches have quickly become very successful in those tasks in audio, often outperforming classical signal processing methods. ML has also swiftly found its way into classification of audio content, e.g., in background-foreground or sound event classification discussed in Section 3.2.1, as well as in other ML-predisposed tasks such as HRTF personalization described in Section 4.2. In DL-driven personalized binaural audio reproduction, a listener can benefit from using individualized HRTFs that are synthesized based on the user's anthropometric features.

We can also observe a rise in popularity of ML in modeling psychoacoustic phenomena. The nonlinear nature of physiology and psychology related to human hearing mechanisms makes ML highly suitable for such tasks. Notable progress has been made, for example, in designing learnable low-level audio features as alternatives to the well-known filterbanks [204–206]. Due to the availability of large amounts of audio content and due to the scarcity of explainable models of the complex human auditory system, we shall expect DL to play an ever increasing role in psychoacoustics in the near future.

Furthermore, multi-modal DL enables spatial audio processing that was either very difficult or even impossible for non-ML-based approaches. One good example is the audio-visual processing covered in Section 3.3, in which multichannel audio is generated from a mono audio signal based on spatial information drawn from the visual content. Since audio-visual dependencies are not straightforward to model mathematically, DL unfolds its potential in finding such complex inter-dependencies, leading to the estimation of, often abstract, representations of spatial audio-visual information.

ML-based approaches have not yet reached a point in which they indisputably surpass in all types of spatial audio processing. One likely reason for the scarcity of powerful DL-based end-to-end models for spatial audio processing is the difficulty to jointly control the timbre, perceptual spatial cues, and audio signal fidelity, as required in high-end applications. Physical accuracy is a criterion that is straightforward to define on a signal level, for example, by means of the squared error. Criteria for achieving a given psychoacoustic result are incomparably more difficult to define because the relation between the signals at the ears of a listener and the resulting perception are known mostly only for relatively simple scenarios [135, 207]. Models for a large range of hearing mechanisms were formulated in [208] from a machine perspective to facilitate integrating them into a machine learning framework. So far, only applications outside of the domain of spatial audio such as music information retrieval have been realized. A proof-of-concept for ML-based assessment of the quality of general (non-spatial) audio was presented in [209], which cannot be generalized to spatial audio. An initial attempt for predicting the perceptual impairment due to system errors in spherical microphone array auralizations using ML for a narrow scope of signals was presented in [210]. Consequently, it still remains a challenge to formulate a single differentiable loss function for neural network training that guarantees that all critically relevant aspects of spatial hearing are represented with the right balance.

Another limiting factor in developing DL-driven spatial audio methods is the lack of available datasets with multichannel audio content at sufficiently large amounts to facilitate the training of DL models. In addition, just the way it is currently unclear how proper loss functions should be defined, the form that ground-truth data should ideally take is also unsettled, which makes data annotation tasks difficult. Until recently, the spatial audio community has not striven to collect and make available multichannel recordings of high fidelity and in large quantity. Note that, in general, the datasets with audio are significantly larger in terms of the data volume than, for example, the datasets for image processing. For instance, the large scale audio-visual dataset of human speech known as VoxCeleb2 [211], which is popular in speaker recognition, contains over one million of speech utterances of a length of 3–20 s with an overall duration of over 2000 h of 16-

bit single-channel audio recorded at a sampling frequency of 16 kHz. This amounts to around 78 GB of data, which is straightforward to handle in practice. In DCASE 2021, the development dataset for the SELD task mentioned in Section 3.2.1 contains overall around 20 hours of 4-channel audio recordings sampled at 24 kHz, which yields already around 13.8 GB of data [107]. In particular, in the case of spatial audio, multiple channels need to be stored for each recording, which poses huge requirements on storage capacity as well as on cache memory and computational power during the training. This is particularly true for more difficult tasks such as those that produce multichannel output data rather than a label, in which the training time can easily last many days even on supercomputers with several GPUs. Most likely for these reasons, only a limited number of datasets with spatial audio, or audio-video, content have been made available.

For instance, binaural audio-visual content is collected in [179, 180, 212], distributed multichannel recordings of multi-speaker conversations are available in [213], while large number of B-format recordings from YouTube are collected in [178]. More such effort and data are required to train generic DL spatial audio models. Furthermore, many available datasets are specific to the input and output setups, i.e., they contain multichannel recordings made using a microphone array with a particular geometry or are synthesized for a particular loudspeaker setup. One remedy to this setup-specific limitation could be to store the captured signals and output signals to be reproduced in one of the commonly accepted representations described in Section 2 such as in the ambisonic format. This way, end-to-end models could be trained irrespective of the capture and reproduction setups. The requirement for standardized capture formats was also identified as an important pre-requisite for databases for training ML systems on the quality of general (non-spatial) audio content [214].

The degree to which the ML-based spatial audio methods will be successful in the future will largely depend on how the audio research community addresses the discussed issues and on the emergence of new applications possible only with DL. For the time being, we are still waiting for the appearance of further disruptive DL-based methods that can deliver spatial audio processing that is out of reach today.

6 Conclusions

We presented an overview of data-based methods in the domain of spatial audio with a special focus on recent approaches that make use of ML. We categorized the methods based on their function in the general signal processing pipeline, which consist of capture, processing, and reproduction.

The capture stage is dominated by solutions that do not employ ML. This is similar for the reproduction stage, in which linear methods are most commmon apart from the task of individualization of the user's head-related transfer functions.

The processing stage is where most of the ML-based solutions are found. For many tasks in this stage, there are both ML-based and non-ML-based methods available. Unlike other domains of data-based processing like visual object recognition and others where the performance of ML-based solutions is significantly superior to non-ML-based ones, such trends have not crystallized in the field of spatial audio. Tasks like source separation, sound event detection, or extraction of spatial information from accompanying video have been highly impacted by DL, leading to outstanding results even with single-channel recordings. Other tasks with a higher focus on extracting or analyzing the spatial properties of sound have not been so much disrupted by ML methods.

Possible causes for this are the lack of robust success criteria (i.e., how to measure that the processing was useful) and, partly as a consequence of this, the amount of the available training and test data, which is lower by orders of magnitude compared to classical application domains of ML.

Abbreviations

6-DoF: 6-Degree-of-freedom; AE: Auto encoder; AR: Augmented reality; BRIR: Binaural room impulse response; CNN: Convolutional neural network; DirAC: Directional audio coding; DL: Deep learning; DNN: Deep neural network; DOA: Direction of arrival; GRNN: Generalized regression neural network; ML: Machine learning; HO: Higher order; HRIR: Head-related impulse response; HRTF: Head-related transfer function; MPEG: Moving Picture Expert Group; NN: Neural network; PCA: Principal component analysis; PWD: Plane wave decomposition; QMF: Quadrature mirror filter; RIR: Room impulse response; RST: Ray space transform; SDM: Spatial decomposition method; SED: Sound event detection; SELD: Sound event detection and localization; SH: Spherical harmonic; SIRR: Spatial impulse response rendering; SMA: Spherical microphone array; SRIR: Spatial room impulse response; VBAP: Vector-base amplitude panning; VR: Virtual reality; WFS: Wave field synthesis

Authors' contributions

All authors contributed equally to this work. All authors read and approved the final manuscript.

Funding

This work received funding from Grant RTI2018-097045-B-C21 funded by MCIN/AEI/10.13039/501100011033 and "ERDF A way of making Europe." Additionally, from the National Science Centre of Poland under grant number DEC-2017/25/B/ST7/01792 and Generalitat Valenciana under grants AICO/2020/154 and AEST/2020/012. Open access funding provided by Chalmers University of Technology.

Availability of data and materials

Not applicable.

Declarations

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 14 June 2021 Accepted: 12 April 2022
Published online: 16 May 2022

References

1. J. Y. Hong, J. He, B. Lam, R. Gupta, W.-S. Gan, Spatial audio for soundscape design: recording and reproduction. Appl. Sci. **7**(6) (2017). https://doi.org/10.3390/app7060627
2. W. Zhang, P. N. Samarasinghe, H. Chen, T. D. Abhayapala, Surround by sound: a review of spatial audio recording and reproduction. Appl. Sci. **7**(5) (2017). https://doi.org/10.3390/app7050532
3. F. Rumsey, Spatial quality evaluation for reproduced sound: terminology, meaning, and a scene-based paradigm. J. Audio Eng. Soc. **50**(9), 651–666 (2002)
4. J. Francombe, T. Brookes, R. Mason, Evaluation of spatial audio reproduction methods (part 1): elicitation of perceptual differences. J. Audio Eng. Soc. **65**(3), 198–211 (2017)
5. M. Cobos, J. J. Lopez, J. M. Navarro, G. Ramos, Subjective quality assessment of multichannel audio accompanied with video in representative broadcasting genres. Multimed. Syst. **21**(4), 363–379 (2015)
6. D. de Vries, in *Second Int. Symp. on Universal Communication*. Wave field synthesis: history, state-of-the-art and future (AES, New York, 2008)
7. J. V. Candy, *Model-based signal processing*. (Wiley-IEEE Press, Hoboken, 2005)
8. A. J. Berkhout, A holographic approach to acoustic control. J. Audio Eng. Soc. **36**(12), 977–995 (1988)
9. V. Pulkki, Virtual sound source positioning using vector base amplitude panning. J. Audio Eng. Soc. **45**(6), 456–466 (1997)
10. U. Horbach, A. Karamustafaoglu, R. Pellegrini, P. Mackensen, G. Theile, in *Audio Engineering Society Convention 106*. Design and applications of a data-based auralization system for surround sound (AES, Munich, 1999)
11. U. Horbach, A. Karamustafaoglu, M. M. Boone, in *Audio Engineering Society Convention 108*. Practical implementation of a data-based wave field reproduction system (AES, Paris, 2000)
12. M. Geier, J. Ahrens, S. Spors, Object-based audio reproduction and the audio scene description format. Organised Sound. **15**(3), 219–227 (2010)
13. P. Annibale, R. Rabenstein, S. Spors, P. Steffen, in *2009 17th European Signal Processing Conference*. A short review of signals and systems for spatial audio (EUSIPCO, Glagow, 2009), pp. 720–724
14. Y. LeCun, Y. Bengio, G. Hinton, Deep learning. Nature. **521**(7553), 436–444 (2015)
15. N. Das, S. Chakraborty, J. Chaki, N. Padhy, N. Dey, Fundamentals, present and future perspectives of speech enhancement. Int. J. Speech Technol. **24**, 1–19 (2020)
16. K. Choi, G. Fazekas, K. Cho, M. Sandler, A tutorial on deep learning for music information retrieval. arXiv preprint arXiv:1709.04396 (2017)
17. J. Blauert, R. Rabenstein, Providing surround sound with loudspeakers: a synopsis of current methods. Arch. Acoust. **37**(1), 5–18 (2012)
18. S. Spors, H. Wierstorf, A. Raake, F. Melchior, M. Frank, F. Zotter, Spatial sound with loudspeakers and its perception: a review of the current state. Proc. IEEE. **101**(9), 1920–1938 (2013). https://doi.org/10.1109/JPROC.2013.2264784
19. J. Blauert, *Spatial hearing: the psychophysics of human sound localization*. (MIT press, Cambridge, 1997)
20. A. C. Keller, Early Hi-Fi and stereo recording at Bell Laboratories (1931-1932). J. Audio Eng. Soc. **29**(4), 274–280 (1981)
21. A. D. Blumlein, Improvements in and relating to sound-transmission, sound-recording and sound-reproducing systems. Patent GB394325A (1933)
22. J. Ahrens, *Analytic methods of sound field synthesis*. (Springer, Heidelberg, 2012)
23. M. Vorländer, *Auralization: fundamentals of acoustics, modelling, simulation, algorithms and acoustic virtual reality*. (Springer, Heidelberg, 2008)
24. B. Rafaely, A. Avni, Interaural cross correlation in a sound field represented by spherical harmonics. JASA. **127**(2), 823–828 (2010)
25. F. Zotter, M. Frank, *Ambisonics: a practical 3D audio theory for recording, studio production, sound reinforcement, and virtual reality*. (Springer, Heidelberg, 2019)
26. J. Ahrens, S. Spors, Wave field synthesis of a sound field described by spherical harmonics expansion coefficients. JASA. **131**(3), 2190–2199 (2012)
27. G. Theile, H. Wittek, M. Reisinger, in *24th Int. Conference of the AES*. Potential wavefield synthesis applications in the multichannel stereophonic world (AES, Banff, 2003)
28. T. Ajdler, L. Sbaiz, M. Vetterli, The plenacoustic function and its sampling. IEEE/ACM Trans. on Sig. Proc. **54**(10), 3790–3804 (2006)
29. R. Mignot, L. Daudet, F. Ollivier, Room reverberation reconstruction: interpolation of the early part using compressed sensing. IEEE/ACM Trans. Audio Speech Lang. Proc. **21**(11), 2301–2312 (2013). https://doi.org/10.1109/TASL.2013.2273662
30. R. Mignot, G. Chardon, L. Daudet, Low frequency interpolation of room impulse responses using compressed sensing. IEEE/ACM Trans. Audio Speech Lang. Proc. **22**(1), 205–216 (2013)
31. N. Antonello, E. De Sena, M. Moonen, P. A. Naylor, T. van Waterschoot, Room impulse response interpolation using a sparse spatio-temporal representation of the sound field. IEEE/ACM Trans. Audio Speech Lang. Proc. **25**(10), 1929–1941 (2017)
32. S. A. Verburg, E. Fernandez-Grande, Reconstruction of the sound field in a room using compressive sensing. JASA. **143**(6), 3770–3779 (2018)
33. F. Katzberg, R. Mazur, M. Maass, P. Koch, A. Mertins, A compressed sensing framework for dynamic sound-field measurements. IEEE/ACM Trans. Audio Speech Lang. Proc. **26**(11), 1962–1975 (2018)
34. S. Emura, in *IEEE Int. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP)*. Sound field estimation using two spherical microphone arrays (ICASSP, New Orleans, 2017), pp. 101–105
35. E. Fernandez-Grande, Sound field reconstruction using a spherical microphone array. J. Acoust. Soc. Am. **139**(3), 1168–1178 (2016)
36. A. M. Torres, J. J. Lopez, B. Pueo, M. Cobos, Room acoustics analysis using circular arrays: an experimental study based on sound field plane-wave decomposition. J. Acoust. Soc. Am. **133**(4), 2146–2156 (2013)
37. E. M. Hulsebos, Auralization using wave field synthesis. Ph. D. Thesis, Delft University of Technology (2004)
38. M. Cobos, S. Spors, J. Ahrens, J. J. Lopez, in *45th Int. AES Conference*. On the use of small microphone arrays for wave field synthesis auralization (AES, Helsinki, 2012)
39. F. Melchior, Investigations on spatial sound design based on measured room impulse responses. PhD thesis, Technische Universität Ilmenau (2011)
40. S. Tervo, J. Pätynen, A. Kuusinen, T. Lokki, Spatial decomposition method for room impulse responses. J. Audio Eng. Soc. **61**(1/2), 17–28 (2013)
41. S. Tervo, J. Pätynen, N. Kaplanis, M. Lydolf, S. Bech, T. Lokki, Spatial analysis and synthesis of car audio system and car cabin acoustics with a compact microphone array. J. Audio Eng. Soc. **63**(11), 914–925 (2015)
42. M. Frank, F. Zotter, in *Proc. of DAGA*. Spatial impression and directional resolution in the reproduction of reverberation (DEGA, Aachen, 2016), pp. 1–4
43. M. Zaunschirm, M. Frank, F. Zotter, Binaural rendering with measured room responses: first-order ambisonic microphone vs. dummy head. Appl. Sci. **10**(5) (2020). https://doi.org/10.3390/app10051631
44. S. A. Garí, J. Arend, P. Calamia, P. Robinson, Optimizations of the spatial decomposition method for binaural reproduction. JAES. **68**(12) (2021). https://doi.org/10.17743/jaes.2020.0063
45. N. A. Gumerov, R. Duraiswami, *Fast multipole methods for the Helmholtz equation in three dimensions*. (Elsevier Science, Amsterdam, 2005), p. 520. https://doi.org/10.1016/B978-0-08-044371-3.X5000-5
46. M. Gerzon, Periphony: with-height sound reproduction. J. Audio Eng. Soc. **21**(1), 2–10 (1973)
47. J. Meyer, G. Elko, in *IEEE International Conference on Acoustics, Speech, and Signal Processing*. A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield (ICASSP, Orlando, 2002), pp. 1781–1784
48. T. D. Abhayapala, D. B. Ward, in *IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2*. Theory and design of high order sound field microphones using spherical microphone array (ICASSP, Orlando, 2002), pp. 1949–1952
49. L. Bianchi, F. Antonacci, A. Sarti, S. Tubaro, The ray space transform: a new framework for wave field processing. IEEE Trans. Signal Process. **64**(21), 5696–5706 (2016). https://doi.org/10.1109/TSP.2016.2591500
50. D. Markovic, G. Sandrini, F. Antonacci, A. Sarti, S. Tubaro, in *IWAENC 2012; International Workshop on Acoustic Signal Enhancement*. Plenacoustic imaging in the ray space (IWAENC, Aachen, 2012), pp. 1–4
51. D. Markovic, F. Antonacci, A. Sarti, S. Tubaro, Soundfield imaging in the ray space. IEEE Trans. Audio Speech Lang. Process. **21**(12), 2493–2505 (2013). https://doi.org/10.1109/TASL.2013.2274697

52. L. Comanducci, F. Borra, P. Bestagini, F. Antonacci, A. Sarti, S. Tubaro, in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. Ray space transform interpolation with convolutional autoencoder (IWAENC, Tokyo, 2018), pp. 261–265. https://doi.org/10.1109/IWAENC.2018.8521397
53. M. Pezzoli, J. J. Carabias-Orti, M. Cobos, F. Antonacci, A. Sarti, Ray-space-based multichannel nonnegative matrix factorization for audio source separation. IEEE Signal Process. Lett. **28**, 369–373 (2021). https://doi.org/10.1109/LSP.2021.3055463
54. M. Pezzoli, F. Borra, F. Antonacci, A. Sarti, S. Tubaro, in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Estimation of the sound field at arbitrary positions in distributed microphone networks based on distributed ray space transform (IEEE, Calgary, 2018), pp. 186–190
55. Z.-Q. Wang, J. Le Roux, J. R. Hershey, in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Multi-channel deep clustering: discriminative spectral and spatial embeddings for speaker-independent speech separation (IEEE, Calgary, 2018), pp. 1–5
56. Z.-Q. Wang, X. Zhang, D. Wang, Robust speaker localization guided by deep learning-based time-frequency masking. IEEE/ACM Trans. Audio Speech Lang. Process. **27**(1), 178–188 (2019). https://doi.org/10.1109/TASLP.2018.2876169
57. C. Knapp, G. Carter, The generalized correlation method for estimation of time delay. IEEE Trans. Acoust. Speech Signal Process. **24**(4), 320–327 (1976). https://doi.org/10.1109/TASSP.1976.1162830
58. E. L. Ferguson, S. B. Williams, C. T. Jin, in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Sound source localization in a multipath environment using convolutional neural networks, (2018), pp. 2386–2390. https://doi.org/10.1109/ICASSP.2018.8462024
59. L. Comanducci, M. Cobos, F. Antonacci, A. Sarti, in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Time difference of arrival estimation from frequency-sliding generalized cross-correlations using convolutional neural networks, (2020), pp. 4945–4949. https://doi.org/10.1109/ICASSP40776.2020.9053429
60. Y. Cao, Q. Kong, T. Iqbal, F. An, W. Wang, M. Plumbley, in *Proc. Detection Classification Acoust. Scenes Events Workshop*. Polyphonic sound event detection and localization using a two-stage strategy (DCASE, New York, 2019), pp. 30–34
61. R. Sato, K. Niwa, K. Kobayashi, Ambisonic signal processing DNNs guaranteeing rotation, scale and time translation equivariance. IEEE/ACM Trans. Audio Speech Lang. Process., 1–1 (2021). https://doi.org/10.1109/TASLP.2021.3069193
62. C. Jin, S. Carlile, Neural system model of human sound localization. Adv. Neural Inf. Process. Syst. **12**, 761–767 (1999)
63. C. Jin, M. Schenkel, S. Carlile, Neural system identification model of human sound localization. J. Acoust. Soc. Am. **108**(3), 1215–1235 (2000). https://doi.org/10.1121/1.1288411
64. Hanaa Mohsin Ali Al- Abboodi, Binaural sound source localization using machine learning with spiking neural networks features extraction. PhD thesis, School of Computing, Science and Engineering, University of Salford-Manchester (2019)
65. E. Thuillier, H. Gamper, I. J. Tashev, in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Spatial audio feature discovery with convolutional neural networks, (2018), pp. 6797–6801. https://doi.org/10.1109/ICASSP.2018.8462315
66. S. K. Zieliński, in *International Conference on Computer Information Systems and Industrial Management*. Improving classification of basic spatial audio scenes in binaural recordings of music by deep learning approach (Springer, Bialystok, 2020), pp. 291–303
67. S. K. Zieliński, H. Lee, P. Antoniuk, O. Dadan, A comparison of human against machine-classification of spatial audio scenes in binaural recordings of music. Appl. Sci. **10**(17), 5956 (2020)
68. J. Vilkamo, T. Bäckström, A. Kuntz, Optimized covariance domain framework for time–frequency processing of spatial audio. J. Audio Eng. Soc. **61**(6), 403–411 (2013)
69. V. Pulkki, U. P. Svensson, Machine-learning-based estimation and rendering of scattering in virtual reality. J. Acoust. Soc. Am. **145**(4), 2664–2676 (2019)
70. Z. Fan, V. Vineet, H. Gamper, N. Raghuvanshi, in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Fast acoustic scattering using convolutional neural networks (IEEE, Barcelona, 2020), pp. 171–175
71. Z. Fan, V. Vineet, C. Lu, T. W. Wu, K. McMullen, in *EEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Prediction of Object Geometry from Acoustic Scattering Using Convolutional Neural Networks, (2021), pp. 471–475. online
72. Z. Tang, H.-Y. Meng, D. Manocha, *Learning Acoustic Scattering Fields for Dynamic Interactive Sound Propagation*, (2021). online
73. R. L. Jenison, A spherical basis function neural network for approximating acoustic scatter. J. Acoust. Soc. Am. **99**(5), 3242–3245 (1996)
74. S. Watanabe, M. Yoneyama, An ultrasonic visual sensor for three-dimensional object recognition using neural networks. IEEE Trans. Robot. Autom. **8**(2), 240–249 (1992)
75. H. Kon, H. Koike, in *Audio Engineering Society Convention 144*. Deep neural networks for cross-modal estimations of acoustic reverberation characteristics from two-dimensional images (AES, Milan, 2018)
76. R. F. Perez, G. Götz, V. Pulkki, in *Proceedings of the 23rd International Congress on Acoustics: Integrating 4th EAA Euroregio, vol. 9*. Machine-learning-based estimation of reverberation time using room geometry for room effect rendering (ICA, Aachen, 2019), p. 13
77. H. Kim, L. Remaggi, P. J. Jackson, A. Hilton, in *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. Immersive spatial audio reproduction for VR/AR using room acoustic modelling from 360 images (IEEE, Osaka, 2019), pp. 120–126
78. H. Kim, L. Remaggi, S. Fowler, P. Jackson, A. Hilton, Acoustic room modelling using 360 stereo cameras. IEEE Trans. Multimedia. **23**, 4117–4130 (2020)
79. C. Schissler, C. Loftin, D. Manocha, Acoustic classification and optimization for multi-modal rendering of real-world scenes. IEEE Trans. Vis. Comput. Graph. **24**(3), 1246–1259 (2017)
80. D. Li, T. R. Langlois, C. Zheng, Scene-aware audio for 360 videos. ACM Trans. Graph. (TOG). **37**(4), 1–12 (2018)
81. Z. Tang, N. J. Bryan, D. Li, T. R. Langlois, D. Manocha, Scene-aware audio rendering via deep acoustic analysis. IEEE Trans. Vis. Comput. Graph. **26**(5), 1991–2001 (2020)
82. F. Lluís, P. Martínez-Nuevo, M. Bo Møller, S. Ewan Shepstone, Sound field reconstruction in rooms: inpainting meets super-resolution. J. Acoust. Soc. Am. **148**(2), 649–659 (2020)
83. O. Ronneberger, P. Fischer, T. Brox, in *International Conference on Medical Image Computing and Computer-assisted Intervention*. U-net: convolutional networks for biomedical image segmentation (Springer, Munich, 2015), pp. 234–241
84. M. S. Kristoffersen, M. B. Møller, P. Martínez-Nuevo, J. Østergaard, Deep sound field reconstruction in real rooms: introducing the isobel sound field dataset. arXiv preprint arXiv:2102.06455 (2021)
85. M. M. J.-A. Simeoni, S. Kashani, P. Hurley, M. Vetterli, Deepwave: a recurrent neural-network for real-time acoustic imaging. Adv. Neural Inf. Process. Syst. 32 (Nips 2019). **32**(CONF), 1–5 (2019)
86. Y. Cai, X. Liu, Y. Xiong, X. Wu, Three-dimensional sound field reconstruction and sound power estimation by stereo vision and beamforming technology. Appl. Sci. **11**(1), 92 (2021)
87. A. S. Bregman, *Auditory scene analysis: the perceptual organization of sound*. (MIT press, Cambridge, 1994)
88. D. Wang, G. J. Brown, *Computational auditory scene analysis: principles, algorithms, and applications*. (Wiley-IEEE press, Hoboken, 2006)
89. M. Brandstein, *Microphone arrays: signal processing techniques and applications*. (Springer, Berlin/Heidelberg, 2001)
90. S. Gannot, E. Vincent, S. Markovich-Golan, A. Ozerov, A consolidated perspective on multimicrophone speech enhancement and source separation. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(4), 692–730 (2017)
91. S. Adavanne, A. Politis, J. Nikunen, T. Virtanen, Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. IEEE J. Sel. Top. Signal Process. **13**(1), 34–48 (2018)
92. S. Chakrabarty, E. A. Habets, Multi-speaker DOA estimation using deep convolutional networks trained with noise signals. IEEE J. Sel. Top. Signal Process. **13**(1), 8–21 (2019)
93. M. J. Bianco, P. Gerstoft, J. Traer, E. Ozanich, M. A. Roch, S. Gannot, C.-A. Deledalle, Machine learning in acoustics: theory and applications. J.

Acoust. Soc. Am. **146**(5), 3590–3628 (2019). https://doi.org/10.1121/1.5133944
94. X. Xiao, S. Watanabe, H. Erdogan, L. Lu, J. Hershey, M. L. Seltzer, G. Chen, Y. Zhang, M. Mandel, D. Yu, in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Deep beamforming networks for multi-channel speech recognition (IEEE, Shanghai, 2016), pp. 5745–5749
95. K. Niwa, T. Nishino, K. Takeda, in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. Encoding large array signals into a 3D sound field representation for selective listening point audio based on blind source separation (IEEE, Las Vegas, 2008), pp. 181–184
96. M. Cobos, J. J. Lopez, Resynthesis of sound scenes on wave-field synthesis from stereo mixtures using sound source separation algorithms. J. Audio Eng. Soc. **57**(3), 91–110 (2009)
97. Q. Liu, W. Wang, P. J. B. Jackson, T. J. Cox, in *2015 23rd European Signal Processing Conference (EUSIPCO)*. A source separation evaluation method in object-based spatial audio, (2015), pp. 1088–1092. https://doi.org/10.1109/EUSIPCO.2015.7362551
98. J. Nikunen, A. Diment, T. Virtanen, M. Vilermo, Binaural rendering of microphone array captures based on source separation. Speech Comm. **76**, 157–169 (2016)
99. Y. Mitsufuji, N. Takamune, S. Koyama, H. Saruwatari, Multichannel blind source separation based on evanescent-region-aware non-negative tensor factorization in spherical harmonic domain. IEEE/ACM Trans. Audio Speech Lang. Process. **29**, 607–617 (2020)
100. Z.-Q. Wang, D. Wang, Combining spectral and spatial features for deep learning based blind speaker separation. IEEE/ACM Trans. Audio Speech Lang. Process. **27**(2), 457–468 (2018)
101. L. Drude, R. Haeb-Umbach, Integration of neural networks and probabilistic spatial models for acoustic blind source separation. IEEE J. Sel. Top. Signal Process. **13**(4), 815–826 (2019)
102. H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, T. Sainath, Deep learning for audio signal processing. IEEE J. Sel. Top. Signal Process. **13**(2), 206–219 (2019). https://doi.org/10.1109/JSTSP.2019.2908700
103. A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, M. D. Plumbley, Detection and classification of acoustic scenes and events: outcome of the DCASE 2016 challenge. IEEE J. Sel. Top. Signal Process. **26**(2), 379–393 (2017)
104. M. C. Green, D. Murphy, in *Proceedings of the Detection and Classification of Acoustic Scenes and Events, Munich, Germany*. Acoustic scene classification using spatial features (DCASE, Munich, 2017), pp. 16–17
105. A. Mesaros, A. Diment, B. Elizalde, T. Heittola, E. Vincent, B. Raj, T. Virtanen, Sound event detection in the dcase 2017 challenge. IEEE/ACM Trans. Audio Speech Lang. Process. **27**(6), 992–1006 (2019)
106. S. Adavanne, P. Pertilä, T. Virtanen, in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Sound event detection using spatial features and convolutional recurrent neural network (IEEE, New Orleans, 2017), pp. 771–775
107. A. Politis, A. Mesaros, S. Adavanne, T. Heittola, T. Virtanen, Overview and evaluation of sound event localization and detection in dcase 2019. IEEE/ACM Trans. Audio Speech Lang. Process. **29**, 684–698 (2020)
108. V. Pulkki, S. Delikaris-Manias, A. Politis, *Parametric time-frequency domain spatial audio*. (Wiley Online Library, Hoboken, 2018)
109. J. Merimaa, V. Pulkki, Spatial impulse response rendering I: analysis and synthesis. J. Audio Eng. Soc. **53**(12), 1115–1127 (2005)
110. V. Pulkki, J. Merimaa, Spatial impulse response rendering II: reproduction of diffuse sound and listening tests. J. Audio Eng. Soc. **54**(1/2), 3–20 (2006)
111. M. Cobos, J. Lopez, S. Spors, A sparsity-based approach to 3D binaural sound synthesis using time-frequency array processing. EURASIP J. Adv. Signal Process. **2010**, 1–13 (2010)
112. V. Pulkki, Spatial sound reproduction with directional audio coding. J. Audio Eng. Soc. **55**(6), 503–516 (2007)
113. G. Del Galdo, M. Taseska, O. Thiergart, J. Ahonen, V. Pulkki, The diffuse sound field in energetic analysis. J. Acoust. Soc. Am. **131**(3), 2141–2151 (2012). https://doi.org/10.1121/1.3682064
114. K. Kowalczyk, O. Thiergart, M. Taseska, G. Del Galdo, V. Pulkki, E. A. P. Habets, Parametric spatial sound processing: a flexible and efficient solution to sound scene acquisition, modification, and reproduction. IEEE Signal Process. Mag. **32**(2), 31–42 (2015). https://doi.org/10.1109/MSP.2014.2369531
115. J. Benesty, C. Jingdong, Y. Huang, *Microphone array signal processing*. (Springer, Berlin, 2008)
116. A. Plinge, S. J. Schlecht, O. Thiergart, T. Robothama, O. Rummukainen, E. Habets, in *AES Int. Conf. on Audio for Virtual and Augmented Reality*. Six-degrees-of-freedom binaural audio reproduction of first-order ambisonics with distance information (AES, Redmond, 2018)
117. M. Kentgens, A. Behler, P. Jax, in *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*. Translation of a higher order ambisonics sound scene based on parametric decomposition, (2020), pp. 151–155. https://doi.org/10.1109/ICASSP40776.2020.9054414
118. J. Vilkamo, T. Lokki, V. Pulkki, Directional audio coding: virtual microphone-based synthesis and subjective evaluation. J. Audio Eng. Soc. **57**(9), 709–724 (2009)
119. A. Politis, M.-V. Laitinen, J. Ahonen, V. Pulkki, Parametric spatial audio processing of spaced microphone array recordings for multichannel reproduction. J. Audio Eng. Soc. **63**(4), 216–227 (2015). https://doi.org/10.17743/jaes.2015.0015
120. K. Kowalczyk, O. Thiergart, A. Craciun, E. A. P. Habets, in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. Sound acquisition in noisy and reverberant environments using virtual microphones, (2013), pp. 1–4. https://doi.org/10.1109/WASPAA.2013.6701869
121. O. Thiergart, G. Del Galdo, M. Taseska, E. A. P. Habets, Geometry-based spatial sound acquisition using distributed microphone arrays. IEEE Trans. Audio Speech Lang. Process. **21**(12), 2583–2594 (2013). https://doi.org/10.1109/TASL.2013.2280210
122. V. Pulkki, M. Karjalainen, *Communication acoustics: an introduction to speech, audio and psychoacoustics*. (Wiley, Hoboken, 2015)
123. O. Thiergart, K. Kowalczyk, E. A. P. Habets, in *2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*. An acoustical zoom based on informed spatial filtering, (2014), pp. 109–113. https://doi.org/10.1109/IWAENC.2014.6953348
124. E. Habets, O. Thiergart, K. Kowalczyk, System, apparatus and method for consistent acoustic scene reproduction based on informed spatial filtering. US Patent 10015613 (2018)
125. A. Favrot, C. Faller, Wiener-based spatial B-format equalization. J. Audio Eng. Soc. **68**(7/8), 488–494 (2020). https://doi.org/10.17743/jaes.2020.0040
126. S. Berge, N. Barrett, in *2nd Int. Symposium on Ambisonics and Spherical Acoustics*. High angular resolution planewave expansion (AmbiSym, Paris, 2010)
127. A. Wabnitz, N. Epain, A. McEwan, C. Jin, in *IEEE Workshop on Appl. of Sig. Proc. to Audio and Acoustics (WASPAA)*. Upscaling ambisonic sound scenes using compressed sensing techniques, (2011), pp. 1–4. https://doi.org/10.1109/ASPAA.2011.6082301
128. L. McCormack, A. Politis, O. Scheuregger, V. Pulkki, in *23rd Int. Congress on Acoustics*. Higher-order processing of spatial impulse responses (ICA, Aachen, 2019)
129. L. McCormack, V. Pulkki, A. Politis, O. Scheuregger, M. Marschall, Higher-order spatial impulse response rendering: investigating the perceived effects of spherical order, dedicated diffuse rendering, and frequency resolution. J. Audio Eng. Soc. **68**(5), 338–354 (2020)
130. A. Politis, J. Vilkamo, V. Pulkki, Sector-based parametric sound field reproduction in the spherical harmonic domain. IEEE J. Sel. Top. Sig. Proc. **9**(5), 852–866 (2015)
131. A. Politis, S. Tervo, V. Pulkki, in *IEEE Int. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP)*. COMPASS: coding and multidirectional parameterization of ambisonic sound scenes, (2018), pp. 6802–6806. https://doi.org/10.1109/ICASSP.2018.8462608
132. W. Oomen, E. Schuijers, B. den Brinker, J. Breebaart, in *Proc. 114th Audio Eng. Soc. (AES) Convention*. Advances in parametric coding for high-quality audio (AES, Milan, 2003)
133. E. Schuijers, J. Breebaart, H. Purnhagen, J. Engdegard, in *Proc. 116th Audio Eng. Soc. (AES) Convention*. Low complexity parametric stereo coding (AES, Berlin, 2004)
134. J. Hilpert, S. Disch, The MPEG surround audio coding standard [standards in a nutshell]. IEEE Signal Proc. Mag. **26**(1), 148–52 (2009). https://doi.org/10.1109/MSP.2008.930433
135. J. Blauert (ed.), *The technology of binaural listening* (Springer, Heidelberg, 2013)
136. F. Baumgarte, C. Faller, Binaural cue coding-part i: psychoacoustic fundamentals and design principles. IEEE Trans. Speech Audio Process. **11**(6), 509–519 (2003). https://doi.org/10.1109/TSA.2003.818109

137. C. Faller, F. Baumgarte, Binaural cue coding-part ii: schemes and applications. IEEE Trans. Speech Audio Process. **11**(6), 520–531 (2003). https://doi.org/10.1109/TSA.2003.818108
138. J. Herre, J. Hilpert, A. Kuntz, J. Plogsties, MPEG-H 3D audio—the new standard for coding of immersive spatial audio. IEEE J. Sel. Top. Signal Process. **9**(5), 770–779 (2015). https://doi.org/10.1109/JSTSP.2015.2411578
139. R. L. Bleidt, D. Sen, A. Niedermeier, B. Czelhan, S. Füg, S. Disch, J. Herre, J. Hilpert, M. Neuendorf, H. Fuchs, J. Issing, A. Murtaza, A. Kuntz, M. Kratschmer, F. Küch, R. Füg, B. Schubert, S. Dick, G. Fuchs, F. Schuh, E. Burdiel, N. Peters, M.-Y. Kim, Development of the MPEG-H TV audio system for ATSC 3.0. IEEE Trans. Broadcast. **63**(1), 202–236 (2017). https://doi.org/10.1109/TBC.2017.2661258
140. J. Herre, H. Purnhagen, J. Koppens, O. Hellmuth, J. Engdegård, J. Hilper, L. Villemoes, L. Terentiv, C. Falch, A. Hölzer, M. L. Valero, B. Resch, H. Mundt, H.-o. Oh, MPEG spatial audio object coding — the ISO/MPEG standard for efficient coding of interactive audio scenes. J. Audio Eng. Soc. **60**(9), 655–673 (2012)
141. P. Coleman, A. Franck, J. Francombe, Q. Liu, T. de Campos, R. J. Hughes, D. Menzies, M. F. S. Gálvez, Y. Tang, J. Woodcock, P. J. B. Jackson, F. Melchior, C. Pike, F. M. Fazi, T. J. Cox, A. Hilton, An audio-visual system for object-based audio: from recording to listening. IEEE Trans. Multimedia. **20**(8), 1919–1931 (2018)
142. Y. Wu, R. Hu, X. Wang, C. Hu, S. Ke, Distortion reduction via cae and densenet mixture network for low bitrate spatial audio object coding. MultiMedia IEEE. **29**(1), 55–64 (2022). https://doi.org/10.1109/MMUL.2022.3142752
143. M. Neuendorf, M. Multrus, N. Rettelbach, G. Fuchs, J. Tobilliard, J. Lecomte, S. Wilde, S. Bayer, S. Disch, C. Helmrich, R. Lefebvre, P. Gournay, B. Bessette, J. Lapierre, K. Kjörling, H. Purnhagen, L. Villemoes, W. Oomen, E. Schuijers, K. Kikuiri, T. Chinen, T. Norimatsu, K. S. Chong, E. Oh, M. Mim, S. Quackenbush, B. Grill, The ISO/MPEG unified speech and audio coding standard — consistent high quality for all content types and at all bit rates. J. Audio Eng. Soc. **61**(12), 956–977 (2013)
144. J. Herre, M. Dietz, MPEG-4 high-efficiency AAC coding [standards in a nutshell]. IEEE Signal Process. Mag. **25**(3), 137–142 (2008). https://doi.org/10.1109/MSP.2008.918684
145. Y. Li, M. Tagliasacchi, O. Rybakov, V. Ungureanu, D. Roblek, in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Real-time speech frequency bandwidth extension, (2021), pp. 691–695. https://doi.org/10.1109/ICASSP39728.2021.9413439
146. A. Biswas, D. Jia, in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Audio codec enhancement with generative adversarial networks, (2020), pp. 356–360. https://doi.org/10.1109/ICASSP40776.2020.9053113
147. N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, M. Tagliasacchi, Soundstream: an end-to-end neural audio codec. IEEE/ACM Transactions on Audio, Speech, and Language Processing. **30**, 495–507 (2022). https://doi.org/10.1109/TASLP.2021.3129994
148. J. Breebaart, C. Faller, *Spatial audio processing: MPEG surround and other applications*. (Wiley, Heidelberg, 2007)
149. C. Avendano, J.-M. Jot, in *Proc. Int. Conf.: Virtual, Synthetic, and Entertainment Audio*. Frequency domain techniques for stereo to multichannel upmix (AES, ESPOO, 2002)
150. C. Uhle, C. Paul, in *Proc. Int. Conf. Digital Audio Effects (DAFx)*. A supervised learning approach to ambience extraction from mono recordings for blind upmixing (DAFx, Helsinki, 2008)
151. K. M. Ibrahim, M. Allam, in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Primary-ambient source separation for upmixing to surround sound systems, (2018), pp. 431–435. https://doi.org/10.1109/ICASSP.2018.8461459
152. S. Y. Park, C. J. Chun, H. K. Kim, in *2016 International Conference on Information and Communication Technology Convergence (ICTC)*. Subband-based upmixing of stereo to 5.1-channel audio signals using deep neural networks, (2016), pp. 377–380. https://doi.org/10.1109/ICTC.2016.7763500
153. K. M. Jeon, S. Y. Park, C. J. Chun, N. I. Park, H. K. Kim, Multi-band approach to deep learning-based artificial stereo extension. ETRI J. **39**(3), 398–405 (2017)
154. J. Choi, J.-H. Chang, Exploiting deep neural networks for two-to-five channel surround decoder. J. Audio Eng. Soc. **68**(12), 938–949 (2021)
155. X. Zheng, Soundfield navigation: separation, compression and transmission. Ph. D. Thesis, University of Wollongong (2013)
156. O. Thiergart, G. D. Galdo, M. Taseska, E. Habets, Geometry-based spatial sound acquisition using distributed microphone arrays. IEEE Trans. Audio Speech Lang. Process. **21**(12), 2583–2594 (2013)
157. C. Schörkhuber, R. Höldrich, F. Zotter, in *Fortschritte der Akustik (DAGA)*. Triplet-based variable-perspective (6DoF) audio rendering from simultaneous surround recordings taken at multiple perspectives, (2020)
158. F. Schultz, S. Spors, in *AES Int. Conf. on Sound Field Control*. Data-based binaural synthesis including rotational and translatory head-movements (AES, Guildford, 2013)
159. Y. Wang, K. Chen, Translations of spherical harmonics expansion coefficients for a sound field using plane wave expansions. JASA. **143**, 3474–3478 (2018)
160. A. Laborie, R. Bruno, S. Montoya, in *114th Conv. of the AES*. A new comprehensive approach of surround sound recording (AES, Amsterdam, 2003)
161. P. Samarasinghe, T. Abhayapala, M. Poletti, Wavefield analysis over large areas using distributed higher order microphones. IEEE/ACM Trans. Audio, Sp. Lang. Proc. **22**(3), 647–658 (2014)
162. N. Ueno, S. Koyama, H. Saruwatari, Sound field recording using distributed microphones based on harmonic analysis of infinite order. IEEE Sig. Proc. Lett. **25**(1), 135–139 (2017)
163. M. Nakanishi, N. Ueno, S. Koyama, H. Saruwatari, in *IEEE Workshop on Appl. of Sig. Proc. to Audio and Acoustics (WASPAA)*. Two-dimensional sound field recording with multiple circular microphone arrays considering multiple scattering (IEEE, New Paltz, 2019), pp. 368–372
164. T. Pihlajamaki, V. Pulkki, Synthesis of complex sound scenes with transformation of recorded spatial sound in virtual reality. JAES. **7/8**(63), 542–551 (2015)
165. K. Wakayama, J. Trevino, H. Takada, S. Sakamoto, Y. Suzuki, in *IEEE Workshop on Appl. of Sig. Proc. to Audio and Acoustics (WASPAA)*. Extended sound field recording using position information of directional sound sources (IEEE, New Paltz, 2017), pp. 185–189
166. L. I. Birnie, T. D. Abhayapala, V. Tourbabin, P. Samarasinghe, Mixed source sound field translation for virtual binaural application with perceptual validation. IEEE/ACM Trans. Audio Speech Lang. Process., 1–1 (2021). https://doi.org/10.1109/TASLP.2021.3061939
167. N. Mariette, B. F. G. Katz, in *EAA Symp. on Auralization*. SoundDelta - largescale, multi-user audio augmented reality (EAA, Espoo, 2009), pp. 1–6
168. E. Bates, H. O'Dwyer, K.-P. Flachsbarth, F. M. Boland, in *144th Conv. of the AES*. A recording technique for 6 degrees of freedom vr (AES, Milan, 2018), p. 10022
169. D. R. Mendez, C. Armstrong, J. Stubbs, M. Stiles, G. Kearney, in *145th Conv. of the AES*. Practical recording techniques for music production with six-degrees of freedom virtual reality (AES, New York, 2018)
170. E. Patricio, A. Ruminski, A. Kuklasinski, L. Januszkiewicz, T. Zernicki, in *Audio Engineering Society Convention 146*. Toward six degrees of freedom audio recording and playback using multiple ambisonics sound fields (AES, Dublin, 2019)
171. J. G. Tylka, E. Y. Choueiri, Domains of practical applicability for parametric interpolation methods for virtual sound field navigation. JAES. **67**(11), 882–893 (2019)
172. K. Müller, F. Zotter, Auralization based on multi-perspective ambisonic room impulse responses. Acta Acustica. **4**(6), 25 (2020). https://doi.org/10.1051/aacus/2020024
173. F. Zotter, M. Frank, C. Schörkhuber, R. Höldrich, in *Fortschritte der Akustik (DAGA)*. Signal-independent approach to variable-perspective (6DoF) audio rendering from simultaneous surround recordings taken at multiple perspectives (DEGA, Hannover, 2020)
174. S. Werner, F. Klein, G. Götz, *Investigation on spatial auditory perception using non-uniform spatial distribution of binaural room impulse responses*, (2019), pp. 137–144. https://doi.org/10.22032/dbt.39967
175. M. Blochberger, F. Zotter, Particle-filter tracking of sounds for frequency-independent 3D audio rendering from distributed B-format recordings. Acta Acustica. **5**, 20 (2021)
176. T. Afouras, A. Owens, J. S. Chung, A. Zisserman, in *16th European Conference on Computer Vision - ECCV, Glasgow, August 23–28*. Self-supervised learning 2070 of audio-visual objects from video, (2020), pp. 208–224

177. R. Gao, K. Grauman, in *Proc. of the IEEE/CVF International Conference on Computer Vision*. Co-separating sounds of visual objects (IEEE, Seoul, 2019), pp. 3879–3888
178. P. Morgado, Y. Li, N. Nvasconcelos. Learning representations from audio-visual spatial alignment, vol. 33, (2020), pp. 4733–4744
179. K. Yang, B. Russell, J. Salamon, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Telling left from right: learning spatial correspondence of sight and sound, (2020), pp. 9929–9938. https://doi.org/10.1109/CVPR42600.2020.00995
180. R. Gao, K. Grauman, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2.5d visual sound, (2019), pp. 324–333. https://doi.org/10.1109/CVPR.2019.00041
181. Y.-D. Lu, H.-Y. Lee, H.-Y. Tseng, M.-H. Yang, in *2019 IEEE International Conference on Image Processing (ICIP)*. Self-supervised audio spatialization with correspondence classifier (IEEE, 2019), pp. 3347–3351
182. A. Rana, C. Ozcinar, A. Smolic, in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Towards generating ambisonics using audio-visual cue for virtual reality (IEEE, Brighton, 2019), pp. 2012–2016
183. H. Huang, M. Solah, D. Li, L.-F. Yu, in *Proceedings of the Conference on Human Factors in Computing Systems*. Audible panorama: automatic spatial audio generation for panorama imagery (CHI, Glasgow, 2019), pp. 1–11
184. S. Paul, Binaural recording technology: a historical review and possible future developments. Acta Acustica U. Acustica. **95**(5), 767–788 (2009)
185. B. Xie, *Head-related transfer function and virtual auditory display*. (J. Ross Publishing, Plantation, 2013)
186. S. Spors, R. Rabenstein, J. Ahrens, in *124th Conv. of the Audio Engineering Society*. The theory of wave field synthesis revisited (AES, Amsterdam, 2008), p. 7358
187. 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 Process. Mag. **34**(3), 36–54 (2017). https://doi.org/10.1109/MSP.2017.2666081
188. D. J. Kistler, F. L. Wightman, A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction. The J. Acoust. Soc. Am. **91**(3), 1637–1647 (1992)
189. M. Zhang, Z. Ge, T. Liu, X. Wu, T. Qu, Modeling of individual hrtfs based on spatial principal component analysis. IEEE/ACM Trans. Audio Speech Lang. Process. **28**, 785–797 (2020)
190. P. Mokhtari, H. Kato, H. Takemoto, R. Nishimura, S. Enomoto, S. Adachi, T. Kitamura, Further observations on a principal components analysis of head-related transfer functions. Sci. Rep. **9**(1), 1–7 (2019)
191. P. Bilinski, J. Ahrens, M. R. P. Thomas, I. J. Tashev, J. Platt, in *IEEE Int. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP)*. HRTF magnitude synthesis via sparse representation of anthropometric features (Florence, Italy, 2014), pp. 4468–4472
192. Y. Shu-Nung, T. Collins, C. Liang, Head-related transfer function selection using neural networks. Arch. Acoust. **42**(3), 365–373 (2017)
193. G. W. Lee, H. K. Kim, Personalized hrtf modeling based on deep neural network using anthropometric measurements and images of the ear. Appl. Sci. **8**(11), 2180 (2018)
194. S. Bharitkar, in *2019 IEEE 9th International Conference on Consumer Electronics (ICCE-Berlin)*. Optimization of head-related transfer function (HRTF) models (IEEE, Berlin, 2019), pp. 251–256
195. T. Chen, T. Kuo, T. Chi, in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Autoencoding HRTFS for DNN based HRTF personalization using anthropometric features, (2019), pp. 271–275. https://doi.org/10.1109/ICASSP.2019.8683814
196. K. Yamamoto, T. Igarashi, Fully perceptual-based 3D spatial sound individualization with an adaptive variational autoencoder. ACM Trans. Graph. (TOG). **36**(6), 1–13 (2017)
197. R. Miccini, S. Spagnol, in *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. HRTF individualization using deep learning, (2020), pp. 390–395. https://doi.org/10.1109/VRW50115.2020.00084
198. S. Spagnol, in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Auditory model based subsetting of head-related transfer function datasets, (2020), pp. 391–395. https://doi.org/10.1109/ICASSP40776.2020.9053360
199. C. Guezenoc, R. Seguier, in *148th AES Convention*. Dataset augmentation and dimensionality reduction of pinna-related transfer functions (AES, Vienna, 2020)
200. B. Rafaely, Analysis and design of spherical microphone arrays. IEEE Trans. Speech Audio Process. **13**(1), 135–143 (2005). https://doi.org/10.1109/TSA.2004.839244
201. F. Brinkmann, A. Lindau, S. Weinzierl, On the authenticity of individual dynamic binaural synthesis. J. Acoust. Soc. Am. **142**(4), 1784–1795 (2017). https://doi.org/10.1121/1.5005606
202. M. Zaunschirm, C. Schörkhuber, R. Höldrich, Binaural rendering of ambisonic signals by head-related impulse response time alignment and a diffuseness constraint. J. Acoust. Soc. Am. **143**(6), 3616–3627 (2018)
203. J. Ahrens, C. Andersson, Perceptual evaluation of headphone auralization of rooms captured with spherical microphone arrays with respect to spaciousness and timbre. J. Acoust. Soc. Am. **145**(4), 2783–2794 (2019). https://doi.org/10.1121/1.5096164
204. M. Ravanelli, Y. Bengio, in *2018 IEEE Spoken Language Technology Workshop (SLT)*. Speaker recognition from raw waveform with sincnet, (2018), pp. 1021–1028. https://doi.org/10.1109/SLT.2018.8639585
205. R. Balestriero, R. Cosentino, H. Glotin, R. Baraniuk, in *Proceedings of International Conference on Machine Learning*. Spline filters for end-to-end deep learning (ICML, Stockholm, 2018), pp. 364–373
206. N. Zeghidour, O. Teboul, F. de Chaumont Quitry, M. Tagliasacchi, in *International Conference on Learning Representations*. LEAF: Aalearnable frontend for audio classification (ICLR, 2021). online
207. J. Blauert, J. Braasch (eds.), *The technology of binaural understanding* (Springer, Heidelberg, 2020)
208. R. F. Lyon, *Human and machine hearing: extracting meaning from sound*. (Cambridge University Press, Cambridge, 2017)
209. C. Volk, J. Nordby, T. Stegenborg-Andersen, N. Zacharov, in *150th Conv. of the Audio Engineering Society*. Predicting audio quality for different assessor types using machine learning (AES, New York, 2021)
210. J. Nowak, G. Fischer, Modeling the perception of system errors in spherical microphone array auralizations. JAES. **67**(12), 994–1002 (2019). https://doi.org/10.17743/jaes.2019.0051
211. J. S. Chung, A. Nagrani, A. Zisserman, in *INTERSPEECH*. Voxceleb2: deep speaker recognition (ISCA, Hyderabad, 2018)
212. S. Wang, A. Mesaros, T. Heittola, T. Virtanen, in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. A curated dataset of urban scenes for audio-visual scene analysis, (2021), pp. 626–630. https://doi.org/10.1109/ICASSP39728.2021.9415085
213. J. Barker, S. Watanabe, E. Vincent, J. Trmal, in *Proc. Interspeech 2018*. The fifth 'CHiME' speech separation and recognition challenge: dataset, task and baselines, (2018), pp. 1561–1565. https://doi.org/10.21437/Interspeech.2018-1768
214. C. Volk, J. Nordby, T. Stegenborg-Andersen, N. Zacharov, in *150th Conv. of the Audio Engineering Society*. Efficient data collection pipeline for machine learning of audio quality (AES, New York, 2021)

Publisher's Note