OPEN ACCESS Freely available online
L. McCormack, A. Politis, T. McKenzie, C. Hold, and V. Pulkki, “Object-Based Six-Degrees-of-Freedom
Rendering of Sound Scenes Captured with Multiple Ambisonic Receivers”
J. Audio Eng. Soc., vol. 70, no. 5, pp. 355–372, (2022 May).
DOI: https://doi.org/10.17743/jaes.2022.0010

Object-Based Six-Degrees-of-Freedom Rendering of Sound Scenes Captured with Multiple Ambisonic Receivers

LEO MCCORMACK,[1] *AES Student Member*, **ARCHONTIS POLITIS,**[2] *AES Associate Member*,
(leo.mccormack@aalto.fi) (archontis.politis@tuni.fi)
THOMAS MCKENZIE,[1] **CHRISTOPH HOLD,**[1] *AES Student Member* **AND VILLE PULKKI,**[1] *AES Fellow*
(thomas.mckenzie@aalto.fi) (christoph.hold@aalto.fi) (ville.pulkki@aalto.fi)

[1]*Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland*
[2]*Department of Information Technology and Communication Sciences, Tampere University, Finland*

This article proposes a system for object-based six-degrees-of-freedom (6DoF) rendering of spatial sound scenes that are captured using a distributed arrangement of multiple Ambisonic receivers. The approach is based on first identifying and tracking the positions of sound sources within the scene, followed by the isolation of their signals through the use of beamformers. These sound objects are subsequently spatialized over the target playback setup, with respect to both the head orientation and position of the listener. The diffuse ambience of the scene is rendered separately by first spatially subtracting the source signals from the receivers located nearest to the listener position. The resultant residual Ambisonic signals are then spatialized, decorrelated, and summed together with suitable interpolation weights. The proposed system is evaluated through an in situ listening test conducted in 6DoF virtual reality, whereby real-world sound sources are compared with the auralization achieved through the proposed rendering method. The results of 15 participants suggest that in comparison to a linear interpolation-based alternative, the proposed object-based approach is perceived as being more realistic.

0 INTRODUCTION

The reproduction of sound scenes captured using a single Ambisonic receiver, for a fixed listening position, is a well-established field. Methods for reproducing Ambisonic signals over a target playback setup may be based either on purely linear mappings [1] or on signal-dependent mappings dictated by (often perceptually motivated) spatial parameters estimated over time and frequency [2]. Both processing approaches may accommodate listener head rotations through either directly rotating the Ambisonic signals [3] or by rotating the directional components prior to spatializing the decomposed scene [4]. Systems that allow for head rotations are said to offer three degrees-of-freedom (3DoF) rendering. Extending the rendering to permit listener translation around a single-receiver is also possible, based on, for example, linear filtering [5] or by employing geometric transformations to appropriately manipulate the directional components used during reproduction [6]. Systems that account for both listener translation and head-rotations are often described as offering six degrees-of-freedom (6DoF), which is a feature that is particularly important for augmented reality (AR) and virtual reality (VR) applications. Single-receiver based translation methods are, however, known to become less robust as the listener moves further away from the capturing point. Multireceiver methods, on the other hand, seek to overcome such limitations by utilizing the additional information afforded by capturing the sound scene from multiple perspectives.

In this work, a system is proposed for the task of using a distributed array of Ambisonic receivers to spatialize recorded sound scenes with 6DoF capability. The system involves the decomposition of the captured sound scene into its individual sound source objects and subsequently spatializing them with respect to the orientation and position of the listener. The system also separately renders the residual ambient components, which represent the Ambisonic receiver signals after the source objects have been subtracted from them. Therefore, in essence, the proposed system may be viewed as a natural multireceiver extension to the Coding and Multi-Parameterisation of Ambisonic Sound Scenes (COMPASS) single-receiver method [7] and is also similar to the approach proposed recently in [8]. With an emphasis on developing a practical system, the proposed processing approach is implemented as a real-time Virtual Studio Technology (VST) audio plug-in[1]. The developed system is then evaluated through subjective listening tests in 6DoF virtual reality, whereby sounds scenes corresponding to a distributed arrangement of seven second-order Ambisonic receivers are rendered over headphones worn by a head-tracked listener and compared directly against the real-life reference scenario and a signal-independent alternative approach [9]. The results suggest that for the majority of tested perceptual attributes and input sound scenes, the proposed method is rated as being closer to the reference compared with the signal-independent baseline approach.

This article is arranged as follows. A literature review of related works is provided in Sec. 1. In Sec. 2, the spatial parameter estimation approaches employed, and the tracking of sound objects in the scene are described. The rendering of both the sources and the ambient components of the scene is then detailed in Sec. 3. Implementation details regarding the developed system are provided in Sec. 4. The subjective listening tests used in the evaluation are described in Sec. 5. The results are then presented and discussed in Sec. 6, along with suggestions for future work. The article is then concluded in Sec. 7.

1 BACKGROUND

In the past, researchers have approached the task of capturing and reproducing sound scenes over an extended spatial region from differing perspectives, targeting different application scenarios with specific requirements. One group of methods requires the use of large loudspeaker arrays, which aim to deliver the appropriate spatial cues to multiple listeners simultaneously over an extended spatial area. Examples of such methods include wave-field synthesis systems [10] for dense 2D loudspeaker distributions or large circular or spherical loudspeaker arrays employing higher-order Ambisonics [11]. However, the capability of such systems to reproduce real spatial recordings remains limited, and therefore, real-world uses of these systems focus on delivery of object-based material. Note that studies related to such systems are not covered in this literature review. Instead, the focus is on approaches that allow the listener to spatially explore recorded sound scenes through translation of the recording point. Such techniques have the potential to serve many simultaneous listeners through individual headphone rendering and are well suited to the rapidly emerging 6DoF immersive media, VR, and AR applications [12].

Proposals that permit the translation of the listening point within a spatial recording may be further categorized as either single-point or multipoint methods, with respect to their recording requirements. Single-point systems, also referred to as sound-field extrapolation methods [13], process a directional recording from a single compact (typically spherical) microphone array, which is able to capture only one perspective of the sound scene. Such methods then aim to extrapolate the rendering to a region beyond the recording center, based on either physical or perceptual considerations [6, 5, 14–20, 13, 21–24]. Because such approaches need only a single multichannel recording, e.g., through the use of an Ambisonic microphone, they are highly efficient in terms of the required equipment and time to set up. However, practical application is typically limited to a small spatial region, since the performance degrades as a function of increased distance between the listener and the recording point [13]. Multipoint methods, on the other hand, utilize recordings from several microphones or microphone arrays, which are distributed over a spatial region of interest. By capturing the same scene from multiple perspectives, such methods permit the rendering of a translated listening point over a larger spatial region [25–38, 9, 39–41, 8, 42–45].

One other distinguishing characteristic between such rendering methods is how they approach the single-point extrapolation or multipoint interpolation process. In this work, a distinction is made between two processing paradigms, termed here as being either nonparametric or parametric. Nonparametric methods do not make assumptions regarding the composition of the sound scene. Rather, they rely on either geometry-informed broad-band remixing and respatialization of the recorded signals [28, 35, 36, 9, 40, 42] or on acoustical sound-field expansions and decompositions, which result in multiple-input–multiple-output (MIMO) filter matrices, to transform the recorded multichannel signals into their translated counterparts from single-point [5, 14, 15, 46, 18, 19, 13, 24] or multipoint [32, 33, 17, 34, 19, 37, 47, 38, 43, 44] recordings. Methods that apply simple broad-band remixing or respatialization of the recorded signals favor implementation simplicity, efficiency, and preservation of the audio fidelity of the original recordings. Some systems have been derived as extensions to surround recording or playback techniques, augmented with listener-position information [35, 36, 40], whereas others perform spatial interpolation directly on distributed Ambisonic recordings, prior to decoding [28, 9, 42]. However, such basic interpolation approaches can result in comb-filtering effects, detrimental coloration, and ambiguous spatial cues during rendering [48].

Other physically inspired nonparametric methods are closely related to the problem of estimating the sound-field over an extended region, based on sampling the spatial pressure or pressure gradients. Their performance therefore depends on the density of this sampling, the wavelength, and the assumed propagation model for sources (far-field, near-field, or mixed). An early precursor of such methods is the work of [49] regarding flexible sound-field recording. Other methods have been based upon plane-wave decompositions [5, 14, 15, 34, 19], point-source decompositions [33], or mixed plane-wave and point-source decompositions [24]. The decompositions in [34, 24] apply additional sparsity constraints. Another popular approach is based on

[1] The real-time system, along with audio examples, may be downloaded from the companion web-page: http://research.spa.aalto.fi/publications/papers/compass_6dof/

the local Fourier–Bessel expansions of the sound field, captured using circular [32, 37, 38] or spherical arrays [18, 43, 44], followed by the re-expansion of the field at a new target position based on a translation operator. Note that the method in [18] requires the additional knowledge of the source position in order to improve the rendering. A comparison of plane-wave decomposition versus re-expansion for single-point recordings is given in [46, 13]. In general, the aforementioned methods achieve high rendering performance mostly at low frequencies or over very limited spatial regions. The multipoint methods based on re-expansion also require a significantly higher number of microphones (or sub-arrays) in order to effectively reconstruct the lower frequencies of the sound field, in a region spanning a couple of meters [37]. The majority of such methods, therefore, tend to present results based only on simulations and do not currently target practical recording scenarios, where broadband performance is desired over a large region based on the input of only a few microphone arrays. Some practical extensions of these approaches are proposed in [48], which integrates information regarding source positions, in order to select subsets of Ambisonic microphones wherein re-expansion works best. This system operates by determining a frequency limit under which said re-expansions are optimal for the given geometry, whereas for higher frequencies, it avoids re-expansions and utilizes a simple spatial interpolation of the Ambisonic signals instead (similar to [9]).

On the other hand, parametric methods for 6DoF rendering employ a spatial model, whereby it is assumed that the composition of the sound scene may be described through spatial parameters. These spatial parameters may be estimated based upon the interchannel dependencies between the multichannel signals of the array and, in the case of multipoint recordings, between the signals of different arrays. Contrary to nonparametric methods, they are typically signal dependent and rely upon the estimation of spatial parameters and the rendering of the audio signals in the time-frequency domain. The assumed models are often the same as those used in parametric spatial audio coding and reproduction methods [2], such as Directional Audio Coding (DirAC) [50] or COMPASS [7]. In these methods, the parameters are utilized for the following: spatial enhancement of the captured scene; flexible rendering beyond linear reproduction capabilities; and/or for spatial modifications of the content. In a listener translation context, the same time-variant parameters, along with information regarding the recording scenario, are used to derive appropriate spatial modifications, which are subsequently applied to the recorded signals to achieve listener translation during rendering. Owing to the additional parameters involved, parametric methods have the potential to achieve effective translation over significantly larger regions, compared with nonparametric, physically inspired methods, provided that the assumed model matches the real scene [47]. Suitable models either assume a single source component with an isotropic diffuse component per time-frequency bin or sub-band [6, 16, 20, 25, 26, 30, 39, 41]; multiple source components accompanied by directional ambience per time-frequency bin [21, 23]; two source components per time-frequency bin [31]; sinusoidal components with spatial noise modeling [29]; or statistically independent source components [27, 31]. Regarding translation of single-point recordings, an early approach, based on DirAC, projected the analyzed DoAs of source signals onto a fixed arbitrary geometry and rendered the source components as point sources while leaving the diffuse component unchanged. This was subsequently explored in a VR and game audio application context [6, 16]. The method was later augmented with known source distance information in [20]. A similar projection approach, based on COMPASS, with a number of multiple time-varying source components, was also studied recently in [21, 23]. The authors of [21] further explore an alternative approach that avoids DoA estimation by applying a primary-ambience separation, projecting the primary component as a distribution of point sources at a fixed radius, and then modifying only this primary component using an Ambisonic warping operation [22] to emulate translation. Another interesting proposal [51] attempts a sparse plane-wave decomposition of the scene, followed by classification and distance estimation of primary and image (reflection) sources, before applying translation operations to them.

The topic of this article falls into the category of multipoint parametric rendering methods. One of the earliest studies in this body of work [25, 26] used widely spaced omnidirectional microphones, filter-bank analysis, stationary versus nonstationary signal decomposition (for foreground/background separation), time-difference of arrival position estimation (for each sub-band), and respatialization of the foreground sources from the analyzed DoAs. Shortly after this study, the work in [27] deployed multiple planar arrays and frequency-domain independent component analysis to separate source components, estimate their positions, and respatialize them for the target listening position. In [29], large planar arrays were used to auralize the interior noise field of an aircraft using sinusoidal components mixed with noise modelling, which was further augmented with spatial covariance matrix modeling. The method was only intended for stationary sound scenes. In [31], a complete parametric pipeline for recordings captured using distributed Ambisonic microphones was presented, including DoA estimation using acoustic intensity vector measurements; source separation between pairs of microphones based on time-frequency masking derived from the DoAs; compression of the separated source signals; and respatialization for translated rendering. In [30], the model of a single-source, plus diffuse component per time-frequency, which was originally exploited in single-point translation in [6, 4], was extended to multipoint recordings. Triangulation of DoAs between arrays permit a point-source position assignment to each time-frequency bin, while a direct-to-diffuse power ratio determines the balance between the source and diffuse signals. This method is also compared against nonparametric methods in [47]; furthermore, the approach is additionally extended in [41], with the simultaneous estimation of source directivities. Recently, [45] used the model of higher-order DirAC [52] to conduct the directional analysis in focused sectors, which were steered towards a particular region of interest, thus reducing interference from out-of-region sources and reverberation.

The closest work to the method proposed in the present article is described in [8], which operated in the time-domain and combined: building 2D planar activity maps based on broadband grid-scanning methods from each receiver, followed by peak-finding to ascertain source position estimates; subsequent particle-filtering based tracking of active sound objects; and then the application of broadband beamforming and spatialization of the objects, mixed with ambient rendering. Building on the work of [8], the proposed system instead operates in the time-frequency domain and lends particular emphasis on real-time operation. It forgoes the use of computationally expensive activity-map–based source position estimation in favor of continuous DoA estimation methods followed by computing the intersecting points between receivers. The proposed method also uses particle-filtering based tracking of sound objects, except this is conducted in 3D space, rather than being restricted to a 2D plane. The proposed method also utilizes the frequency resolution of the employed transform domain to apply spatial post-filters onto the source object signals. The post-filters can suppress leakage of diffuse noise and interference into the broadband beamformed signals when the target source is not active or at frequencies not populated by the source spectrum when it is active.

Regarding perceptual evaluations of 6DoF rendering systems, only a handful of the studies mentioned in this section conducted formal listening tests. Those that did followed different test design philosophies in order to assess specific aspects of the systems. They also permitted varying levels of listener navigable freedoms. This is largely due to the difficulty of conducting such listening tests, since a comprehensive perceptual evaluation would require a real-time dynamic implementation of the system, along with, for example, a head-mounted display (HMD) providing visual context for the audio rendering, while simultaneously presenting an interactive test interface to the listener [20]. Therefore, nonparametric methods based on sound-field decompositions or expansions have primarily relied on comparing pressure reconstruction errors, which may be used to instead infer their perceptual performance [37, 32]. The use of physical proxy measures for the perceived localization and envelopment performance, such as the intensity and energy vectors and diffuseness metrics, was investigated in [46]. Alternatively, the use of auditory modeling was explored in [14] in order to assess localization performance for single-point translation based on a plane-wave decomposition. The single-point parametric method of [51] was also evaluated objectively based on a speech quality metric, which was used to assess the performance at static listener positions in a room containing up to three speakers.

Evaluations involving actual listening tests were conducted for the systems described in [26, 27, 31, 16, 20, 9, 21, 42, 24, 8]. The mean opinion scores for the multipoint source separation method of [27] indicated robust spatial reproduction compared with a stereo reference but lower naturalness compared with a mono reference. The single-point DirAC–based method of [16] showed high perceptual quality in a fixed-listener multiple-stimulus test when compared with reference signals rendered using a receiver at the true translated positions. A similar system was tested in [20], except using dynamic real-time rendering with a test interface in VR. The results indicated significant improvements using known distance information of the sources when compared with the fixed projection radius imposed in [16]. Similar results were reported in [21], while providing additional insights regarding the perceptual effects of using either known or estimated source DoAs and demonstrating degraded performance for large translation distances. In [9, 42], nonparametric spatial interpolation of multiple ambisonic recordings was evaluated with dynamic rendering and aimed to assess the perceived spatial impression and naturalness. The tests showed improvements when increasing the ambisonic order from first to third, given a reference derived from an object-based representation of the same scene involving a few distinct sources [9]. These improvements were largely diminished, however, when interpolating between the many sources of a classical orchestra in a concert hall setting [42]. A nonparametric sound-field extrapolation method was evaluated in [24] using dynamic rendering of a single free-field speech/music source. The work compares mixed and sparse sound-field decomposition approaches against conventional plane-wave nonsparse alternatives, with demonstrably better results when using the former. Finally, listening tests of the system described in [8] were conducted using either fixed listener positions or a dynamic listener following a predefined linear trajectory. It was demonstrated that the system outperformed nonparametric interpolation based alternatives in the majority of cases.

Finally, although not the focus of this work, it is noted that a closely related and equally active topic of research involves interpolation, extrapolation, or parameterization of spatial room impulse responses, which are subsequently re-rendered at a new point. This, therefore, has strong potential in applications such as interactive auralization, VR, and AR. Indeed, in this field, some of the models for single-point or multipoint translated rendering (both parametric and nonparametric), described in the aforementioned treatise, have also found application here. However, spatial RIRs are of short finite-length and exhibit a special structure that allows modeling and processing methods that are not suitable for audio recordings, while their processing is typically more robust to waveform errors or other processing artefacts. This related literature is too large to be covered in the present report; however, the reader is directed to [53] for a recent overview of such techniques.

2 SPATIAL ANALYSIS

The objective of the spatial analysis is to determine the Cartesian coordinates of all active source objects within the scene, taking into account that their number and relative positions may change over time. In this work, a narrow-band, single-source assumption is imposed onto each receiver, and the requisite time-frequency indices, which correspond to a dominant sound source, are identified and used to obtain the corresponding DoA estimates. Light rays are then cast outwards from the receiver positions in the directions of their respective DoA estimates, with the ray intersections in 3D space then found and subsequently passed to a particle-filtering–based tracker. The function of the tracker is to follow clusters of intersection points over time, with the center points of clusters assumed to correspond to source objects within the scene. The proposed rendering aims to then use these tracked sound objects to synthesize a perceptually plausible auralization of the sound scene, from the perspective of an arbitrary listener position, as described later in Sec. 3.

2.1 The Preliminaries

It is assumed that R Ambisonic receivers of arbitrary spherical harmonic order $N_1, ..., N_R$ are either distributed along a line ($R \geq 2$), or on a 2D plane ($R \geq 3$), or alternatively define an arbitrary 3D volume ($R \geq 4$). The signals for each receiver are denoted as $\mathbf{x}_1(t, f), ..., \mathbf{x}_R(t, f) \in \mathbb{C}^{(N_R+1)^2 \times 1}$, which are represented in the time-frequency domain, where t and f denote the down-sampled time and frequency indices, respectively. The Cartesian coordinates for each receiver are denoted as $\mathbf{u}_1(t), ..., \mathbf{u}_R(t) \in \mathbb{R}^{3\times 1}$. Note that the Ambisonic channel numbering (ACN) and ortho-normalized (N3D) Ambisonics conventions are followed throughout this work. The second-order statistics associated with each receiver are represented by their spatial covariance matrices (SCM)

$$\mathbf{C}_{\mathbf{x},r}(t, f) = \mathcal{E}[\mathbf{x}_r(t, f)\mathbf{x}_r^{\mathrm{H}}(t, f)], \quad \text{for } r = 1, ..., R, \tag{1}$$

where $\mathcal{E}[.]$ denotes the expectation operator, which, in practice, often involves temporal averaging in the range of tens of milliseconds. This temporal averaging influences the performance and the responsiveness of the spatial parameter estimation, but it also helps alleviate problems arising due to misaligned receiver positions (with respect to their specified positions), analog-to-digital converter clock synchronization drift offsets, and cases in which the source-to-receiver distances vary across receivers. Due to this latter point, if the receivers are spaced further apart, temporal averaging should also be increased, so that the SCMs of the receivers may still sufficiently encapsulate active source signal statistics during each spatial analysis frame. Note that the time and frequency indices are henceforth omitted for brevity of notation.

Much of the spatial analysis is then based upon the subspace decomposition of the receiver SCMs, which, with the single-source assumption, is given as

$$\begin{aligned}\mathbf{C}_{\mathbf{x},r} &= \mathbf{V}_r \Sigma_r \mathbf{V}_r^{\mathrm{H}}, \\ &= \sigma_{1,r}\mathbf{v}_{1,r}\mathbf{v}_{1,r}^{\mathrm{H}} + \sum_{k=2}^{(N_r+1)^2} \sigma_{k,r}\mathbf{v}_{k,r}\mathbf{v}_{k,r}^{\mathrm{H}}, \\ &\quad \text{for } r = 1, ..., R,\end{aligned} \tag{2}$$

where σ are the eigenvalues sorted in descending order, and $\mathbf{v}$ are the respective eigenvectors. Note that, in the single-source case, the eigenvector corresponding to the largest eigenvalue $\sigma_{1,r}$ is referred to as the signal subspace, whereas the eigenvectors corresponding to the smallest $(N_r + 1)^2 - 1$ eigenvalues are collectively referred to as the noise subspace. In this work, it is assumed that each time-frequency index will correspond to either to a single dominant directional source, expressed by the signal subspace or to weak interferers and ambient noise, as described by the noise subspace.

2.2 Source Signal Detection

The first step of the spatial analysis is to detect the number of active sources from the perspective of each receiver and for each frequency band. There are a number of approaches that have been proposed for this task, with many of them operating based upon the aforementioned subspace decomposition of the SCM; these include: those based on the spatial covariance matrix eigenvalues with thresholding, eigenvalue statistics, or those which operate directly on the eigenvectors. A review of such methods may be found in [54]. However, these methods may generally err on the side of being more *permissive*; in other words, they can have the tendency to overestimate the true number of sources. Although such estimators have been shown to lead to perceptually robust parametric reproduction in [7]; for this work, a single-source assumption was selected instead, in order to better isolate only the dominant sound sources in the scene. One approach for ascertaining which time-frequency tiles correspond only to a dominant source is to combine diffuseness parameter estimation with the application of appropriate thresholding. Examples of diffuseness estimators include those based on active-intensity [2] or on the variance of the eigenvalues [55]. In this work, however, the direct-path-dominance (DPD) test was selected due to its simplicity and robust performance [56]. The DPD test is based on determining the ratio between the largest and second largest eigenvalues as

$$\psi_r = \begin{cases} 1, & \text{if } \frac{\sigma_{1,r}}{\sigma_{2,r}} \geq \lambda \\ 0, & \text{otherwise} \end{cases}, \tag{3}$$

where λ is a threshold value, which is typically tuned so that a small percentage (for example: $\approx 10\%$) of time-frequency tiles pass the DPD test. When $\psi_r = 1$, it is assumed that the time-frequency tile for receiver r comprises only a single dominant source, and thus, the DoA of the source is subsequently estimated as described in the following section. Whereas $\psi_r = 0$, on the contrary, indicates the presence of multiple sources and/or diffuse noise, and therefore, no DoA estimate is made.

2.3 Source Direction Estimation

One popular approach for estimating the DoA of a sound source, from the perspective of a receiver, is to first create an activity map. These may be generated by scanning a dense grid of directions on the unit sphere, using conventional beamformers, followed by the computation of their respective powers. Suitable beamformers for this task include the hyper-cardioid (maximum directivity) [57] or the minimum-variance distortionless response (MVDR) [58] beamformers. Alternatively, a spatial pseudo-spectrum may be computed based on, for example, the MUtiple-SIgnal Classification (MUSIC) approach [59] or the min-norm method [60], both of which operate based upon the noise subspace of the SCM. The DoA estimate, expressed as a unit length Cartesian vector $\boldsymbol{\gamma}_r \in \mathbb{R}^{3\times 1}$, may then be determined as the direction that maximizes/minimizes (method dependent) the generated activity map.

One issue with grid-based DoA estimation methods, however, is the high computational requirements needed for an exhaustive search of the sphere, which is then compounded by the need to analyze the scene from multiple perspectives, as in the present multireceiver case. Furthermore, although grid densities of approximately 1,000 directions have been employed for single-perspective rendering and shown to be perceptually sufficient in [7, 61], it is noted that quantization errors are inherently introduced. When subsequently computing the intersecting points of DoA estimates, such quantization errors may result in large deviations between the estimated source positions and the true source positions, especially if the sources are located far away from the receivers. Therefore, high-precision gridless DoA estimation alternatives may be preferred for multireceiver applications, even if their accuracy is potentially lower, since they mitigate quantization errors and are generally less computationally demanding.

One example of a gridless method is to infer that the opposite direction to the active-intensity vector corresponds to the DoA [62, 63], which is an approach used by some existing parametric reproduction methods [50, 64]. This active-intensity based approach is limited to single-source DoA estimation and first-order input. In the present case, the single-source assumption is not a limitation; however, it may be beneficial to make use of higher-order components, if they are available. Although there are higher-order extensions to the active-intensity DoA estimation approach [65], these employ the higher orders to spatially partition the sphere into individual sectors so that multiple sources may be resolved simultaneously, thus offering minimal benefit when employing the current single-source assumption. Therefore, in order to target a general solution, the EigenBeam Estimation of Signal Parameters via Rotational Invariance Techniques (EB-ESPRIT) approach [66, 67] was selected for this study; since it is: gridless, extends well to higher-order input (if available), and the formulations described in [66, 67] have been specifically shown to be robust in [68]. Furthermore, although EB-ESPRIT is inherently a multisource estimator, simplified formulations of EB-ESPRIT for single-source localization tasks [67] may also be used to further reduce the computational complexity.

2.4 Finding the Intersecting Points

Once the DoAs of dominant sound sources have been estimated from the perspective of each receiver, the proposed system subsequently casts out the corresponding *light rays* from the receiver positions. This process is conducted independently for each frequency band. For all combinations of receiver pairs, which are both adjacent to each other and mutually agree on the presence of a single dominant source, the intersection point is found between the cast rays. In the presence of noise, however, the rays will rarely perfectly intersect in 3D space. Therefore, the mid-point between the shortest distance between the cast rays is considered to be an intersection point in this work. For example, if receivers 1 and 2 mutually agree on the presence of a single source and are adjacent to each other, the intersection is computed as

$$\mathbf{p} = \frac{1}{2}\left(\mathbf{u}_1 + \tau_1\boldsymbol{\gamma}_1 + \mathbf{u}_2 + \tau_2\boldsymbol{\gamma}_2\right), \tag{4}$$

with

$$\tau_1 = \frac{(\mathbf{u}_2-\mathbf{u}_1)^{\mathrm{T}}\boldsymbol{\gamma}_1+(\mathbf{u}_1-\mathbf{u}_2)^{\mathrm{T}}\boldsymbol{\gamma}_2(\boldsymbol{\gamma}_1^{\mathrm{T}}\boldsymbol{\gamma}_2)}{1-(\boldsymbol{\gamma}_1^{\mathrm{T}}\boldsymbol{\gamma}_2)^2}, \tag{5}$$

$$\tau_2 = \frac{(\mathbf{u}_1-\mathbf{u}_2)^{\mathrm{T}}\boldsymbol{\gamma}_2+(\mathbf{u}_2-\mathbf{u}_1)^{\mathrm{T}}\boldsymbol{\gamma}_1(\boldsymbol{\gamma}_1^{\mathrm{T}}\boldsymbol{\gamma}_2)}{1-(\boldsymbol{\gamma}_1^{\mathrm{T}}\boldsymbol{\gamma}_2)^2}. \tag{6}$$

Note that Eq. (4) only provides valid intersection points if $\tau_1 > 0$ and $\tau_2 > 0$. Furthermore, additional heuristics may also be integrated into the system at this point, in order to assess the validity of the calculated intersections. For example, if the length of the shortest distance between the cast rays exceeds a maximum threshold (e.g., 30 cm), then the DoA estimates may be deemed inaccurate, and this intersection may be discarded. Other heuristics may make use of the known layout of the scene, for example: if intersections occur too near to a receiver, then they may be assumed to be erroneous. Another example is where the geometry of the room is known, and therefore, intersection points that fall beyond the room boundaries may be assumed to correspond to early reflections or noisy DoA estimates.

All D narrow-band intersections deemed to be valid, based on the aforementioned heuristics criteria, are then gathered for each analysis frame: $\mathbf{p}_1, ..., \mathbf{p}_D$, where $\mathbf{p}_d \in \mathbb{R}^{3\times 1}$ are the Cartesian coordinates for intersection d. These intersection points should ideally form clusters around the true source positions and be subsequently followed over time by the tracker described in the next section.

2.5 Source Position Tracking

Once all valid DoA intersections have been determined, the data are subsequently passed onto the Rao–Blackwellised particle-filter tracking framework described in [69]. The purpose of the employed tracker is to use the computed intersection points (which are determined from narrow-band single-source DoA estimates) to derive K broadband source position estimates, which are updated for every analysis frame and are denoted as $\mathbf{v}_1, ..., \mathbf{v}_K$, where $\mathbf{v}_k \in \mathbb{R}^{3\times 1}$ are the Cartesian coordinates for source object k.

The tracker models the source object dynamics using the Wiener velocity model [70], in order to adapt to moving sources. The *death* of currently tracked source objects is modeled based on their respective *lifespans* (which are assumed to follow a Gamma distribution) and is considered at every time step. The *birth* of a new source object is postulated by the tracker for each new intersection point it receives and is carried out independently by each particle. The likelihood of this event occurring is weighed-up against the following alternative event hypotheses: (1) the intersection point instead corresponds to an existing tracked source object or (2) the intersection point is deemed to be *clutter*. Each particle randomly selects one of the event hypotheses, which are weighted based upon the estimated likelihood of their occurrence. Particles that consistently select likely event hypotheses over time are subsequently weighted higher and therefore contribute the most to the tracking results. Owing to the described birth and death modeling, the tracking framework is able to follow multiple sound objects, which may vary in number and/or position over time. To better suit the present application, the employed tracking framework also imposes a maximum threshold for the number of targets, which serves to constrain the tracker in cases where its parameters may be poorly tuned in practice. If two targets are located within a certain specified distance, an additionally added feature forces the probability of death for the less mature of the two targets to 1. Note that the employed tracking framework is described in further detail in [71, 72, 69].

3 SPATIAL SYNTHESIS

3.1 Source Stream Rendering

Once the positions of source objects in the scene are being successfully tracked, a dedicated rendering stage is used to spatialize them from the perspective of the listener position over the target playback setup. Here, the nearest receiver to each source is first determined, since it is assumed that beamforming from these receivers will likely provide the highest signal-to-noise (SNR) ratio. The receiver-to-source directions for the K source beamformers are therefore computed as

$$\boldsymbol{\gamma}_{r_k,k} = \frac{\mathbf{v}_k - \mathbf{u}_{r_k}}{||\mathbf{v}_k - \mathbf{u}_{r_k}||}, \quad \text{for } k = 1, ..., K, \tag{7}$$

where $\|.\|$ denotes the Euclidean norm and r_k is the index of the nearest receiver to source object k. The source object signals are then given as

$$s_k = P(\boldsymbol{\gamma}_{r_k,k})\mathbf{w}^{\mathrm{H}}(\boldsymbol{\gamma}_{r_k,k})\mathbf{x}_{r_k}, \quad \text{for } k = 1, ..., K, \tag{8}$$

where $\mathbf{w} \in \mathbb{C}^{(N_{r_k}+1)^2\times 1}$ are beamforming weights, and $P \in \mathbb{R}$ is a frequency-dependent spatial post-filter gain factor. Note that the beamforming weights may, for example, be derived based on frequency-independent axisymmetric patterns [73], or frequency-dependent adaptive algorithms, such as the MVDR [58] beamformer design.

Note that the intention of the included spatial post-filter is to frequency-dependently deactivate the beamformers and mitigate problems arising during the following three potential scenarios: (1) during periods when the source is not currently active but is still being tracked; (2) when the source does not have energy in all frequency bands, and thus, the broad-band nature of the tracker and subsequent beamforming results in the capture of unwanted signals/noise in those frequency bands; and (3) situations in which the system is erroneously tracking a phantom target. In this work, a post-filter is constructed based on the cross-pattern coherence (CroPaC) [74] between an omnidirectional signal and a dipole beamformer that has its positive-polarity lobe steered towards the source direction. The post-filter may be formulated as

$$P(\boldsymbol{\gamma}_{r_k,k}) = \max\left[\lambda, \frac{\sqrt{\frac{1}{3}}\mathcal{R}[\mathbf{C}^{(rot)}_{\mathbf{x}_{r_k}}(\boldsymbol{\gamma}_{r_k,k})]_{1,4}}{\frac{1}{4}\sum_{q=1}^{4}[\mathbf{C}_{\mathbf{x}_{r_k}}]_{q,q}}\right], \tag{9}$$

where $\mathcal{R}[.]$ denotes the real operator and $\mathbf{C}^{(rot)}_{\mathbf{x}_{r_k}}(\boldsymbol{\gamma}_{r_k,k}) = \mathbf{R}(\boldsymbol{\gamma}_{r_k,k})\mathbf{C}_{\mathbf{x}_{r_k}}\mathbf{R}^{\mathrm{H}}(\boldsymbol{\gamma}_{r_k,k})$ are the signal statistics corresponding to an appropriately rotated sound field, (using the azimuth/elevation angles of $\boldsymbol{\gamma}_{r_k,k}$ as yaw/pitch rotations), which aligns the x-axis dipole with the source direction using rotation matrix $\mathbf{R} \in \mathbb{R}^{(N_{r_k}+1)^2\times(N_{r_k}+1)^2}$ [3]. The normalized coherence values are therefore in the range [−1, 1] and become unity when the dipole is steered toward an active source signal. The $\lambda > 0$ parameter is then used to half-wave rectify and constrain the coherence values between [λ, 1], in order to ensure that the resulting post-filter is not influenced by the negative-polarity lobe of the dipole beamformer and to also help retain the signal fidelity of the beamformer. Note that when $\lambda = 0$, the post-filter is permitted to completely attenuate time-frequency tiles, and this freedom can lead to the introduction of spectral artefacts, whereas, for example, $\lambda = 0.25$, would instead permit attenuation of up to approximately 12 dB, which may largely mitigate such issues.

Once estimates of the source object signals have been obtained, they are then spatialized with respect to the listener position $\mathbf{a}_l \in \mathbb{R}^{3\times 1}$, which is also expressed as Cartesian coordinates, in meters. The listener-to-source direction is calculated as

$$\boldsymbol{\gamma}_{l,k} = \frac{\mathbf{v}_k - \mathbf{a}_l}{||\mathbf{v}_k - \mathbf{a}_l||}, \quad \text{for } k = 1, ..., K, \tag{10}$$

and the spatialization for an S-channel playback setup may be realized with spatialization gains $\mathbf{g}(\boldsymbol{\gamma}_{l,k}) = [g_1(\boldsymbol{\gamma}_{l,k}), ..., g_S(\boldsymbol{\gamma}_{l,k})]^{\mathrm{H}}$, which can be, for example, HRTFs for binaural playback, arbitrary-order spherical harmonic weights for Ambisonics output, or amplitude-panning gains for loudspeaker playback. The spatialization is then applied as

$$\mathbf{y}_{\mathrm{dir}} = \sum_{k=1}^{K} \frac{d_{r_k,k}}{d_{l,k}}\mathbf{g}(\boldsymbol{\gamma}_{l,k})s_k, \tag{11}$$

where $d_{r_k,k}$ and $d_{l,k}$ are the distances between the (nearest) receiver-to-source and listener-to-source, respectively, and are included in order to account for distance attenuation according to the inverse-distance law. Note that due to the narrow-band processing paradigm employed, frequency-dependent distance filters could also be feasibly integrated into the system, in order to account for near-field/proximity effects [75, 76], although these were not explored in this present study. Furthermore, unlike linear interpolation-based rendering, it is worth highlighting that this object-based source stream rendering is also able to support listener positions that fall outside the area/volume enclosed by the receiver positions.

3.2 Ambient Stream Rendering

The purpose of the ambient stream rendering is to reproduce diffuse ambient components and weak directional sources, which remain after the source components have been subtracted from the input Ambisonic signals. Firstly, the nearest J receivers to the listener are identified, where the number is dependent on the receiver arrangement; for example, the nearest $J = 2$ receivers are selected for a line array arrangement, $J = 3$ for the 2D planar case, and $J = 4$ for the general 3D volumetric case. The indices $r_1, ..., r_J$ are subsequently used to obtain the residual Ambisonic signals as [7]

$$\begin{aligned}\hat{\mathbf{x}}_{r_j} &= \mathbf{x}_{r_j} - \mathbf{Y}(\boldsymbol{\Gamma}_{r_j})\mathbf{P}(\boldsymbol{\Gamma}_{r_j})\mathbf{W}(\boldsymbol{\Gamma}_{r_j})\mathbf{x}_{r_j} \\ &\quad \text{for } j = 1, ..., J, \end{aligned} \tag{12}$$

where $\mathbf{W} = (\mathbf{Y}^{\mathrm{T}}\mathbf{Y})^{-1}\mathbf{Y}^{\mathrm{T}} \in \mathbb{C}^{K\times(N_{r_j}+1)^2}$ are beamforming weights for all source directions $\boldsymbol{\Gamma}_{r_j} = [\boldsymbol{\gamma}_{r_j,1}, ..., \boldsymbol{\gamma}_{r_j,K}]$ from the perspective of the jth nearest receiver (r_j); $\mathbf{P} = \mathrm{diag}[P(\boldsymbol{\gamma}_{r_j,1}), ..., P(\boldsymbol{\gamma}_{r_j,K})]$ is a diagonal matrix of CroPaC spatial post-filter gains; and $\mathbf{Y} \in \mathbb{R}^{(N_{r_j}+1)^2\times K}$ are spherical harmonic re-encoding weights for the same directions as used for the beamforming. Note that during periods of source inactivity, where the post-filter can introduce significant attenuation, this processing can lead to $\hat{\mathbf{x}}_{r_j} \approx \mathbf{x}_{r_j}$. Furthermore, in order to stabilize the pseudo-inverse, when determining $\mathbf{W}$ in practice, DoAs that fall within the same $\pi/(2N_r)$ angular window may be replaced by a single averaged DoA vector.

A plane-wave decomposition of the residual receiver signals is then conducted, followed by scaling these signals with linear or barycentric interpolation gains $b_1, ..., b_J$, before accumulating and spatializing the resultant signals as

$$\mathbf{y}_{\mathrm{diff}} = \mathbf{G}_{\mathrm{v}} \textstyle\sum_{j=1}^{J} b_j \mathcal{D}[\mathbf{Y}_{\mathrm{v}}^{\mathrm{T}}\hat{\mathbf{x}}_{r_j}], \tag{13}$$

where $\mathbf{Y}_{\mathrm{v}} \in \mathbb{R}^{V\times(N_{r_j}+1)^2}$ are spherical harmonic weights for a uniform spherical arrangement of V (*virtual loudspeaker*) directions. The gain b_j dictates the degree to which each receiver contributes to the residual rendering (based on the chosen interpolation scheme), and $\mathbf{G}_{\mathrm{v}} \in \mathbb{C}^{S\times V}$ are spatialization gains to map the signals corresponding to the virtual directions to that of the target playback setup. Since the encapsulated ambient components may be assumed to be mostly diffuse, timbral colorations (such as position-dependent comb-filtering, which results from the coherent summation of signals and is experienced with linear interpolation rendering alternatives) are likely to be avoided. However, in order to enforce this diffuse property, the decomposed signals may also be decorrelated, as denoted by $\mathcal{D}[.]$, before they are included in the weighted average.

Note that, if the listener is located away from a receiver position, then the intention of this ambient rendering is not to be physically accurate but rather to produce a plausible rendition of the diffuse ambience of the scene, whereas when the listener is located at one of the receiver positions, then the rendering reverts to that of the physically motivated ambient rendering conducted by the methods described in [7, 61]. Furthermore, unlike the situation with the source stream rendering, the proposed ambient rendering does not support extrapolation beyond the convex hull of the receiver positions. In such cases, the listener position for the ambient rendering may instead be *pegged* to the nearest point on the convex hull of the receiver positions. However, it is noted that the overall rendering may remain perceptually plausible, since this particular limitation may be masked by the extrapolation capabilities of the source stream rendering.

3.3 Overall Rendering

The final output parametrically rendered signals are obtained by summing the two streams as

$$\mathbf{y}_{\mathrm{par}} = \mathbf{y}_{\mathrm{dir}} + \mathbf{y}_{\mathrm{diff}}, \tag{14}$$

where it is apparent that one could optionally apply different gains to the two streams, in order to either emphasize the source object rendering (which would be akin to dereverberation) or emphasize the reverberance of the scene. Additionally, linearly interpolated and rendered output signals may be obtained as

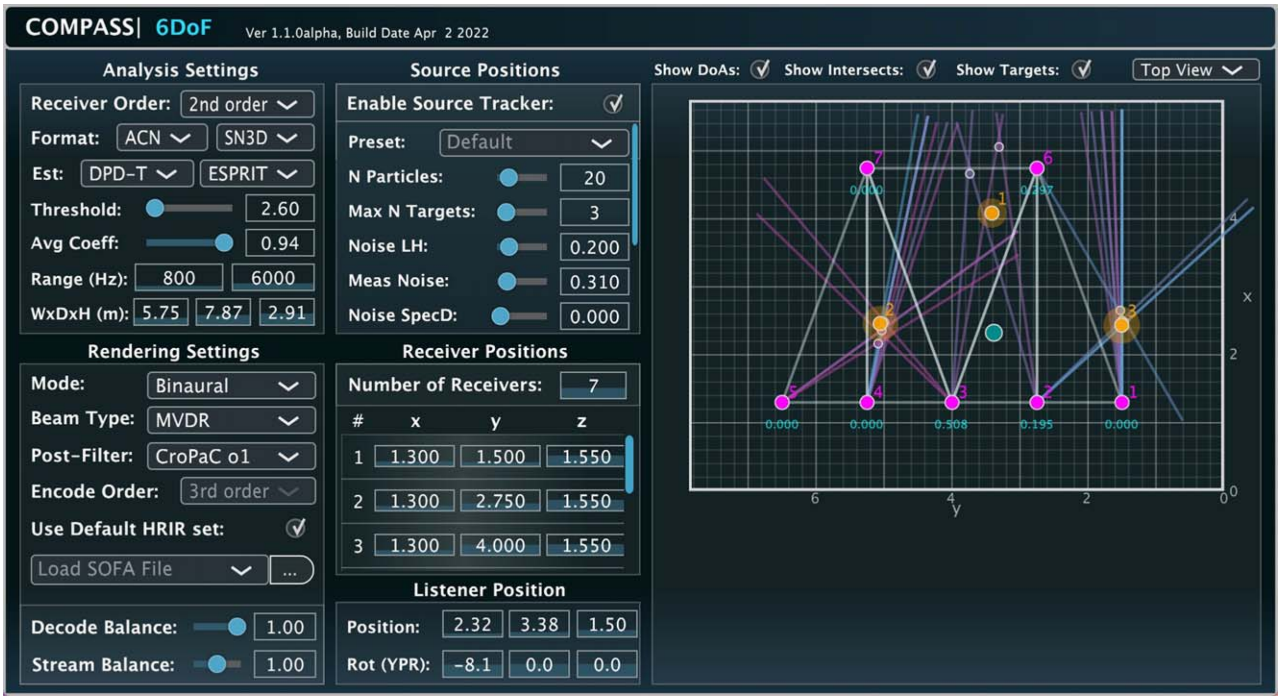
$$\mathbf{y}_{\mathrm{lin}} = \sum_{j=1}^{J} b_j \mathbf{D}_{\mathrm{lin},r_j} \mathbf{x}_{r_j}, \tag{15}$$

where $\mathbf{D}_{\mathrm{lin},r_j} \in \mathbb{C}^{S\times(N_{r_j}+1)^2}$ are static spatialization gains for linearly mapping the interpolated Ambisonic signals to the target playback format; for example, a binaural Ambisonic decoder [77, 78] may be used for headphone playback. Although it is expected that this linear rendering will not be as spatially accurate as the proposed approach, the output signals may exhibit higher signal fidelity; since no signal-dependent processing is conducted. A weighted combination of the proposed parametric rendering and this optional linear rendering may therefore be beneficial in some scenarios. Note that these nonparametrically rendered signals also serve as the baseline approach during the evaluations described in Sec. 5.

Finally, it is noted that additional listeners, each with their own position and orientation, may experience a personalized reproduction of the sound scene with minimal extra computational requirements; since the proposed spatial analysis described in Sec. 2 will remain the same regardless of the listener positions. Furthermore, the beamformer signals will also remain the same, with only the spatialization gains and delay-compensation required to be updated for each listener. The residual signals may also be reused across listeners, provided that their respective interpolation gains are applied accordingly.

4 IMPLEMENTATION

The proposed multireceiver object-based sound scene reproduction system was implemented as a real-time VST audio plugin, using the open-source Spa-



COMPASS| 6DoF Ver 1.1.0alpha, Build Date Apr 2 2022

Analysis Settings

Receiver Order: 2nd order
Format: ACN SN3D
Est: DPD-T ESPRIT
Threshold: 2.60
Avg Coeff: 0.94
Range (Hz): 800 6000
WxDxH (m): 5.75 7.87 2.91

Rendering Settings

Mode: Binaural
Beam Type: MVDR
Post-Filter: CroPaC o1
Encode Order: 3rd order
Use Default HRIR set: ✓
Load SOFA File ...
Decode Balance: 1.00
Stream Balance: 1.00

Source Positions

Enable Source Tracker: ✓
Preset: Default
N Particles: 20
Max N Targets: 3
Noise LH: 0.200
Meas Noise: 0.310
Noise SpecD: 0.000

Receiver Positions

Number of Receivers: 7

#	x	y	z
1	1.300	1.500	1.550
2	1.300	2.750	1.550
3	1.300	4.000	1.550

Listener Position

Position: 2.32 3.38 1.50
Rot (YPR): -8.1 0.0 0.0

Show DoAs: ✓ Show Intersects: ✓ Show Targets: ✓ Top View

Fig. 1. The graphical user interface of the developed Virtual Studio Technology (VST) audio plug-in.

tial_Audio_Framework[2] and the JUCE framework.[3] The graphical user interface is depicted in Fig. 1, which is divided into six sections: analysis and rendering settings; source, receiver, and listener positions; and a window providing a visual depiction of the multireceiver setup. The receiver positions are specified to lie within the boundaries of a shoe-box room geometry, which serves to constrain the spatial analysis to operate within a known volume by culling intersections that fall beyond these boundaries. It is possible to also support arbitrary room geometries, although this is not explored in this work. Note, however, that constraining the problem within a known geometry is also not a strict requirement, and if intersections located beyond the boundaries are not culled, then the room viewing window serves only to depict the relative positions of the receivers and sound objects.

Once the receiver positions (depicted as magenta-colored circles) have been specified, their input signals are transformed into the time-frequency domain using the alias-free short-time Fourier transform (STFT) design described in [79], which was configured with a hop size of 128 samples and 90% overlap. The audio latency of the system at a 48-kHz sample rate is therefore 24 ms. However, since the listener position and orientation are updated after the forward time-frequency transform, the head-tracking latency incurred by the synthesis stage is instead $13.\dot{3}$ ms, whereas the latency for the spatial parameter estimation is dependent on the amount of temporal averaging applied to the SCMs ($\geq 13.\dot{3}$ ms). The SCMs are also grouped and averaged into equivalent rectangular bandwidths (ERB) frequency bands. This averaging serves to improve the robustness of the spatial parameter estimation (without having to resort to longer temporal averaging and thus incurring more latency), while also reducing the computational complexity of the real-time system.

The spatial analysis is conducted for each Ambisonic receiver and ERB frequency band using the DPD test and EB-ESPRIT methods described in Secs. 2.2 and 2.3, respectively. The source position estimates are subsequently found by casting rays based on the DoA estimates from adjacent receiver pairs and computing the points of intersection, if the receivers mutually agree on the presence of a single-dominant source, as described in Sec. 2.4. These cast DoA rays and intersecting points are also depicted on the user interface with a color gradient ranging from magenta to cyan, in order to indicate whether the estimates/intersections correspond to lower- or higher-frequency bands, respectively. Intersections that fulfill the validity criteria (i.e., are within the room boundaries, not closer than 30 cm to a receiver, and the distance between intersections and rays is within 30 cm) are given to the multisource particle-filtering based tracker described in Sec. 2.5. The tracked source positions are then depicted as orange circles, with their positional variance indicated by transparent halos. It is then based upon these tracked source positions that the rendering operates, unless the tracker is disabled and the source positions are specified manually instead.

After the source positions have been specified or tracked, the system may be configured with one of the three following options: (1) to output both the source and ambient streams corresponding to the specified listener position (depicted as a green colored circle) and orientation, as described in Sec. 3, using HRTFs as the spatialization gains; (2) to render both streams using spherical harmonic vec-

[2] https://github.com/leomccormack/Spatial_Audio_Framework
[3] https://github.com/juce-framework/JUCE

tors as the spatialization gains instead, in which case any existing Ambisonic decoder may then be used to auralize the translated scene; or (3) to render only the source stream without spatialization and output the individual source object signals (one per output channel), along with exposing their positions as read-only automation/metadata, which may be subsequently spatialized using a separate tool. The proposed system is therefore highly flexible, and as far as the authors are aware, is the only multireceiver object-based rendering system that is publicly available. Furthermore, it is noted that since the layout of the receivers may be re-configured (and associated parameters reinitialized) on a frame-by-frame basis, the system can not only support moving sound sources and listeners but also moving receivers. However, moving sources and/or receivers were not formally investigated in this study.

For the source beamforming, a few static and adaptive algorithm options were integrated into the system, with the minimum-variance distortionless response (MVDR) beamformer design [58] chosen for this present study. A variety of different post-filtering options were explored by the authors, but the first-order CroPaC algorithm ($\lambda = 0.25$) was selected due to its robust performance [74] and low computational complexity. The residual stream signals are decorrelated through a combination of applying short frequency-dependent delays [80] and cascaded all-pass filters [81], as commonly conducted by multichannel audio codecs [82]. Barycentric interpolation weights are employed for the ambient stream interpolation according to Eq. (13), and the interpolation-based rendering alternative described by Eq. (15).

Time-delay compensation is applied by the system for each source object signal, in order to account for the difference between receiver-to-source and listener-to-source distances. To preserve a causal processing behavior, a fixed delay of 15 ms is added by the system to allow for up to 8.5 m of listener translation from the sound sources. Note also that the compensation is applied in the time-frequency domain without fractional-delay based interpolation and, therefore, time delays are quantized to the nearest down-sampled time index. Due to the large STFT overlap in the proposed configuration, however, this lack of fractional delays was not found to noticeably affect the perceived performance, and its omission significantly reduced the computational complexity of the system.

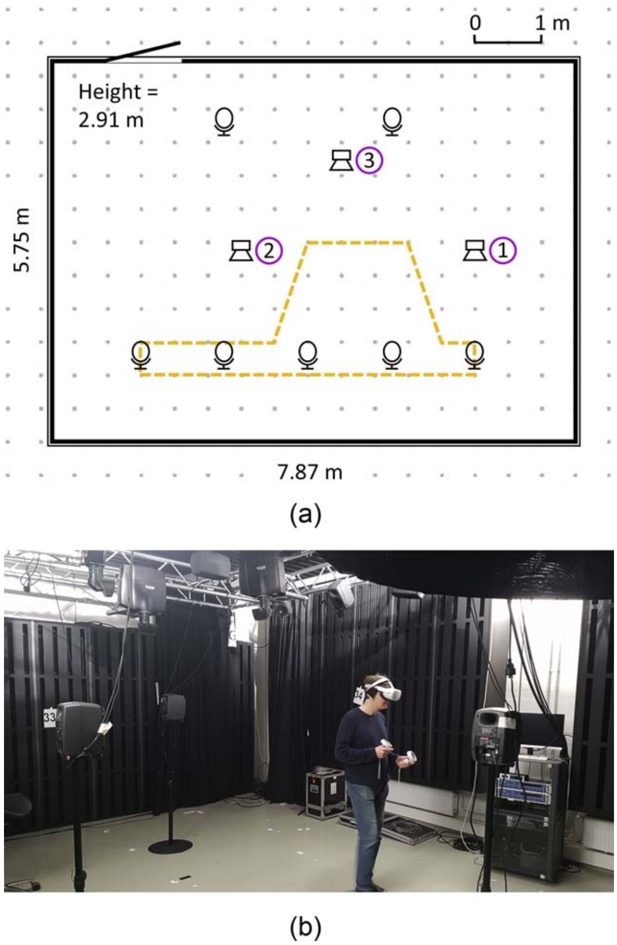


Fig. 2. (a) Illustration of the receiver and source positions used during the SRIR measurements and the listening test, with the permitted navigable area indicated by the yellow dashed line. (b) Photo of the room and loudspeaker positions, and a listener located within the navigable area.

5 EVALUATION

The evaluation of the proposed system was approached through an in situ subjective listening test conducted in 6DoF virtual reality, whereby real-world source signals were directly compared with the corresponding binaural renderings of a distributed arrangement of Ambisonic receivers. The dataset described in [83][4] was employed for this evaluation, which comprises spatial room impulse responses (SRIRs) measured in the 5.75 × 7.87 × 2.91 m (width × depth × height) Arni variable acoustics room at the Acoustics Lab, Aalto University, Finland (background noise level of 20.5 dBA SPL). The dataset involved placing an mh Acoustics Eigenmike em32 in seven different receiver/measurement positions and capturing the SRIRs for three different source positions, which were represented by Genelec 8331A coaxial loudspeakers. The source and receiver positions are illustrated in Fig. 2. Note that the SRIRs measured during the dry room configuration, with octave-band (125 Hz to 8 kHz) RT60s of [0.23 0.22 0.28 0.30 0.50 0.46 0.36] s, were selected for the present study.

A multiple-stimulus listening test was designed, whereby loudspeakers of the same make and model were placed in the same source positions in the room. This therefore, served as the real-world reference (*ref*) test condition. The other three test conditions then comprised (1) the parametric system using known source positions (i.e., tracker disabled) in conjunction with the proposed spatial synthesis (*par*); (2) the parametric system using both the proposed spatial analysis (i.e., tracker enabled) and spatial synthesis (*parT*); and (3) a linear baseline approach (*lin*), as described in [9] and represented by Eq. (15). The test conditions are

[4] https://doi.org/10.5281/zenodo.5720724

Table 1. Listening test conditions.

Name	Processing method
ref	Stimuli played directly over loudspeakers
par	Proposed system (known source positions)
parT	Proposed system (with tracker enabled)
lin	Linear interpolation and decoding

Table 2. Listening test scene stimuli, ordered according to the source numbering in Fig. 2.

Name	Source stimuli
speech	male speech, female speech, male speech
mix	male speech, piano, water fountain
band	drums, bass guitar, strings

also summarized in Table 1. Regarding the stimuli used for the three source positions, three different combinations of anechoic monophonic recordings were selected: (1) three simultaneous speakers (*speech*); (2) a mixture of speech, piano, and a water fountain (*mix*); and (3) drums, bass guitar, and strings (*band*). The stimuli combinations are summarized in Table 2.

For the reference (and hidden-reference test case), the stimuli were simply played directly through the three loudspeakers in the room. For the other test conditions, the Eigenmike SRIRs were first encoded into second-order Ambisonics and subsequently convolved with the same stimuli for each of the seven receiver positions. The resulting synthetic distributed microphone array recordings were then rendered binaurally over headphones through the application of the developed VST plug-in, which was able to switch between the three processing methods under test. The test participants were therefore able to directly compare the three processing methods (and hidden-reference) against the reference case. To display the listening test room in virtual reality, a three-dimensional model of the same room was also captured using light detection and ranging (LIDAR) technology using an Apple iPad Pro. The model was refined and reduced in file size in Blender, to improve real-time rendering performance, before being imported into Unity. Note that the loudspeaker positions were also visible in the room model.

To maximize the available computational resources, two computers were used to run the listening test. The virtual reality visuals were rendered with Unity using a Windows laptop, whereas the developed VST plug-in and the listening test logic were hosted within Cycling 74 Max using a MacOS laptop. User position and orientation data for listener translation and rotation, as well as the controls for the graphical user interface for the test, were sent from Unity to Max via open sound control (OSC) messages. To prevent participants from colliding with the loudspeakers, an appropriate navigable area was determined and displayed as a barrier within the virtual reality environment; this area is also illustrated in Fig. 2(a). The visuals were delivered over an Oculus Quest 2 head-mounted display, and the auralization was delivered using Mysphere 3.2 headphones. The Oculus Quest 2 communicated with Unity wirelessly over Oculus Air Link, whereas the headphone connection was wired.

The test participants were fitted with the head-mounted display and headphones and were requested to familiarize themselves with the virtual reality environment and user interface prior to beginning the listening test. The duration of the test was between 30 and 45 min, and although no training phase was included, participants could take as long as they wished. The test subjects were also encouraged to walk freely inside the navigable area and were requested to look and move around before making their assessments. The participants were required to separately rate the conditions based on their *Spatial* and *Timbral* similarity with the reference and then also based upon their *Overall* preference. The test scenes and processing conditions were randomized and double blind. Trials were repeated once; therefore, given the three sets of stimuli, there were six trials in total. The 15 participants, all of whom were either employees or postgraduate students at the Acoustics Lab, were aged between 24 to 35 (12 male, three female) with self-reported normal hearing and prior critical listening experience. Note that a REAPER project, which includes the seven second-order Ambisonic receiver signals employed and rendered by the developed VST audio plug-in, may also be found on the companion webpage.[1]

6 RESULTS AND DISCUSSION

The results of the listening test are presented as violin plots in Fig. 3, which shows density trace and a box plot in a single illustration, thus depicting the structure of the data in greater detail than a traditional box plot [84]. The width of the violins indicates the density of data, and median values are presented as a white circle. The interquartile range is marked using a thick grey line, the range between the lower and upper adjacent values is marked using a thin grey line, and individual results are displayed as colored dots. Through high-level observations, it is noted that the reference was consistently identified and rated with the highest scores, whereas the median values for the two parametric approaches are situated between the reference and the linear method for all tested stimuli types and evaluation metrics. The results also appear to indicate that the two parametric processing approaches were rated similarly but that the spatial analysis conducted by *parT* introduces more variance in the results and therefore, these cases were rated slightly lower than *par*, which operated with known source positions. However, the median scores for both parametric approaches are still shown to be higher than those attained with the linear baseline approach.

In order to gain further insight into the results, the data were tested for normality using the Shapiro–Wilk test, which showed 12 out of 36 results data followed a nonnormal distribution (at a confidence interval of 95% p). Statistical analysis was therefore conducted using nonparametric methods. To assess whether the results of the rendering methods being studied were statistically significantly different, Friedman tests were conducted for the *Spatial* and

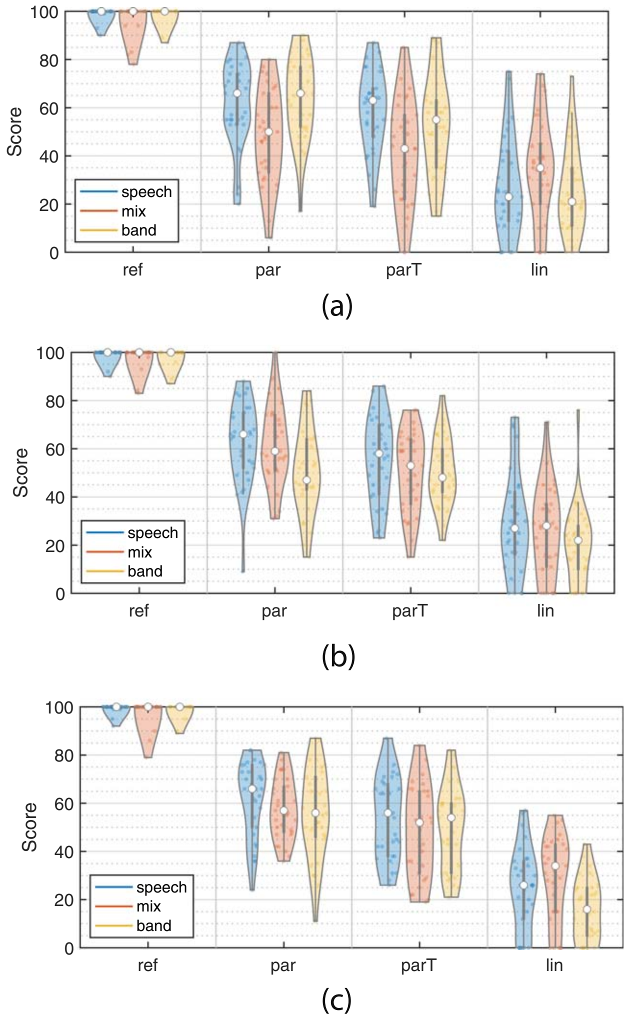


Fig. 3. Violin plots of the listening test results for the three evaluation metrics under test: (a) *Spatial* similarity, (b) *Timbral* similarity, and (c) *Overall* preference.

Timbral evaluation metric results, for all three stimuli types. The results were found to be significant for *Spatial*: $\chi^2(3) = 80.1, p < 0.01$, $\chi^2(3) = 64.5, p < 0.01$ and $\chi^2(3) = 81.6$, $p < 0.01$ for the *speech*, *mix*, and *band* stimuli, respectively. They were also significant for *Timbral*: $\chi^2(3) = 75.9$, $p < 0.01$, $\chi^2(3) = 82.1, p < 0.01$ and $\chi^2(3) = 82.9, p < 0.01$ for the *speech*, *mix*, and *band*, respectively. Since the *Overall* part of the listening test was deemed to be more subjective and to represent a perceptual average of the *Spatial* and *Timbral* metrics, further analysis of the *Overall* results was not conducted in this study.

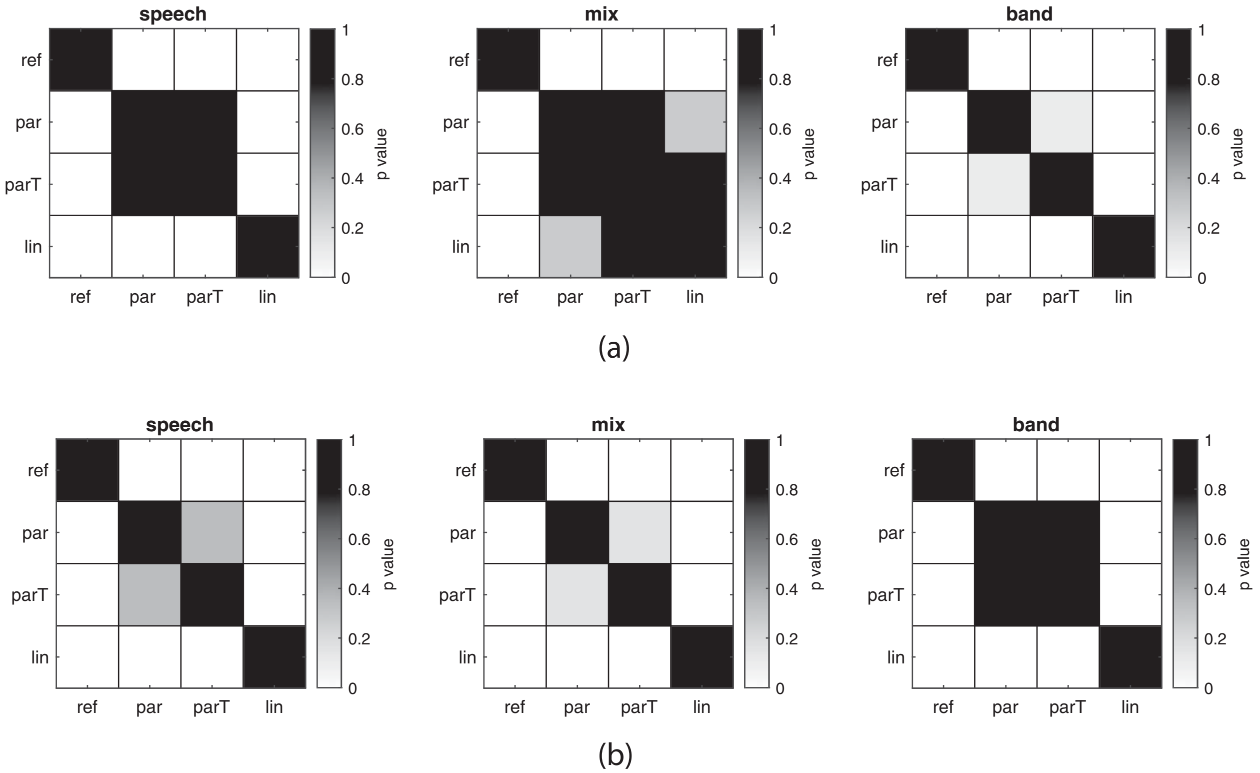
To explore the significance of the differences between rendering methods in more detail, pairwise post-hoc Wilcoxon Signed Rank tests were conducted between all conditions for the *Spatial* and *Timbral* metrics, with the Bonferroni–Holm p value correction, the results of which are presented in Fig. 4. In all tested scenarios, the hidden reference was rated statistically significantly higher than all other conditions, suggesting that none of the tested rendering methods were spatially or timbrally transparent with reality. However, for the *Timbral* metric, the parametric rendering approaches were both rated statistically significantly higher than the linear rendering for all three stimuli types. For the *Spatial* metric, this was significant for the *speech* and *band* stimuli but not the *mix* stimulus. No statistically significant differences were observed between the two parametric rendering methods for either the *Spatial* or *Timbral* results, for all tested stimuli types. This confirms that rendering the scene using the source tracker produces perceptually similar outcomes to when rendering using known source positions.

6.1 Avenues for Future Work

During the course of the study, it became apparent that attaining a physically accurate rendering, in such a way that it would be transparent with respect to the real-life reference, poses a significant challenge. Therefore, the study instead targeted a perceptually plausible rendering of the 6DoF scenario, rather than a transparent rendering. This is due in part to some practical limitations that are not easy, or even possible, to overcome. For example, the HRTFs employed by the rendering methods under test were non-individualized, and therefore, the performance may vary across listeners. However, perhaps more significant, is the fact that while the Mysphere 3.2 headphones employed have been shown to be more acoustically transparent than other commonly used headphone models for binaural tests [85], the headphones and head-mounted display geometry still incur scattering and occlusion effects. These effects will introduce direction-dependent filtering when auditioning real-life reference conditions. Investigating ways to improve the acoustical transparency of the testing apparatus is likely to be an important topic when conducting similar perceptual studies in the future.

One of the main limitations of the proposed rendering is the absence of source directivity modeling. Although it is acknowledged that detailed modeling of source directivity would likely require a large number of receivers-to-source ratio, it is nonetheless highlighted as an important research direction for future systems that target a more physically accurate rendering. One possible simplification of this task would be to assume that the source directivity of all the sources within the scene is the same (e.g., having a directivity pattern which is standardized to that of a human speaker or loudspeaker), in which case, the problem would be simplified and reduced to only finding the source orientations. The source orientation could be approximated based upon beamforming from a few nearby receivers, from which the ratio of energy between the low- and high-frequencies could be computed; following the assumption that a speaker or loudspeaker would be more directive at high frequencies than at low frequencies, an estimate of the source orientation could subsequently be established. This also extends to modeling near-field effects, which may improve plausibility when in close proximity to a sound source.

Regarding the implementation of the proposed system, it is noted that a primary reason for selecting seven second-order receivers as input for the evaluation (63 channels in total) was to stay within the 64-channel limit imposed by



(a)

(b)

Fig. 4. Wilcoxon signed-rank matrices of the *Spatial* and *Coloration* listening test results between different conditions: (a) *Spatial* similarity, and (b) *Timbral* similarity.

the VST2 Source Development Kit (SDK). Future systems based on more and higher-order receivers will likely require custom software solutions or will need to distribute multiple instances of the plug-in over different buses. Furthermore, it is noted that one of the more significant computational burdens of the present system is the need for the decorrelation of multiple plane-wave decomposed receiver signals in the ambient stream rendering, prior to summation and spatialization in Eq. (13). However, if the system is configured for binaural playback, then theoretically, only two channels of decorrelated audio should be required to produce the appropriate interchannel relationships corresponding to a diffuse-field, as demonstrated recently in [86]. Therefore, to reduce the computational complexity, similar solutions could be explored for the proposed ambient rendering.

7 CONCLUSION

This article proposes a practical system for object-based rendering of sound scenes captured using a distributed arrangement of Ambisonic receivers. The proposed system operates in the time-frequency domain and employs sound source direction-of-arrival (DoA) estimates from the perspective of each receiver, followed by tracking the resulting clusters of DoA intersection points. The employed tracker can adapt to sound sources which vary in number and position over time, with broad-band beamformers and narrow-band spatial post-filters used to extract their respective signals. The post-filters are intended to frequency-dependently deactivate the beamformers during periods of source inactivity, or when the sources themselves excite only a limited frequency range. The extracted source object signals are then spatialised with respect to the orientation and position of the listener, who does not necessarily need to be located within the area/volume enclosed by the receiver positions. Additionally, the source object signals are subtracted from the receivers closest to the listener, where the resulting residual signals are subsequently spatialised, decorrelated, and summed with appropriate interpolation weights; which therefore enables rendering of also the diffuse reverberant sound components of the captured sound scene.

To evaluate the perceptual performance of the proposed rendering techniques, they were first integrated into a real-time software solution. A subjective listening test experiment was then designed, whereby simulated recordings of seven second-order Ambisonic receivers were generated based on three different source positions in a real room. Listening test subjects were then placed in the same room, where real loudspeakers were situated in the same three source positions as used to generate the simulated recordings. The participants were also provided with acoustically transparent headphones, and an HMD depicting the loudspeaker positions in a virtual representation of the room. This testing apparatus, therefore, allowed the participants to make a direct comparison between the real-life situation and the respective binaural renderings of the simulated signals, using the methods under test, which were: the proposed rendering using sound source tracking, the proposed rendering using known source positions instead, and a signal-independent interpolation-based alternative.

The results show that in the vast majority of cases, the proposed renderings—both with the tracker and using known source positions—were rated to be perceptually closer to the reference, compared with the linear interpolation baseline alternative. No statistically significant differences were found between cases when the source object tracker was enabled or disabled, suggesting that the proposed tracking was perceptually adequate for the evaluated scenes. Furthermore, the reference was consistently rated statistically significantly higher than the proposed renderings, which suggests that the system was not perceptually transparent with respect to reality. The authors postulate a number of aspects of the proposed system and evaluation apparatus, which may be investigated in future studies; but chiefly, the lack of source directivity, and near-field effect modeling and the direction-dependent scattering introduced by the headphones and HMD, are highlighted as the most significant focus areas for future work.

8 ACKNOWLEDGMENTS

This research has received funding from the Aalto University Doctoral School of Electrical Engineering. Thomas McKenzie was supported by the Human Optimised XR (HumOR) Project. Special thanks are extended to Nils Meyer-Kahlen for assistance with the 3D room scan and Unity scene used during the evaluation and to Sebastian J. Schlecht for the insightful discussions that transpired during the course of this study.

9 REFERENCES

[1] F. Zotter and M. Frank, *Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality* (Springer Nature, Berlin, Germany, 2019). https://doi.org/10.1007/978-3-030-17207-7.

[2] V. Pulkki, S. Delikaris-Manias, and A. Politis, *Parametric Time-Frequency Domain Spatial Audio* (John Wiley & Sons, Hoboken, NJ, 2017).

[3] J. Ivanic and K. Ruedenberg, "Rotation Matrices for Real Spherical Harmonics. Direct Determination by Recursion," *J. Phys. Chem. A*, vol. 102, no. 45, pp. 9099–9100 (1998 Nov.). https://doi.org/10.1021/jp953350u.

[4] A. Politis, T. Pihlajamäki, and V. Pulkki, "Parametric Spatial Audio Effects," in *Proceedings of the 15th International Conference on Digital Audio Effects (DAFx-12)* (York, UK) (2012 Sep.).

[5] F. Schultz and S. Spors, "Data-Based Binaural Synthesis Including Rotational and Translatory Head-Movements," in *Proceedings of the 52nd International Conference of AES: Sound Field Control - Engineering and Perception* (2013 Sep.), paper P-7.

[6] T. Pihlajamäki and V. Pulkki, "Projecting Simulated or Recorded Spatial Sound onto 3D-Surfaces," in *Proceedings of the 45th International Conference of the AES: Applications of Time-Frequency Processing in Audio* (2012 Mar.), paper 4-5.

[7] A. Politis, S. Tervo, and V. Pulkki, "COMPASS: Coding and Multidirectional Parameterization of Ambisonic Sound Scenes," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6802–6806 (2018 Apr.). https://doi.org/10.1109/ICASSP.2018.8462608.

[8] M. Blochberger and F. Zotter, "Particle-Filter Tracking of Sounds for Frequency-Independent 3D Audio Rendering from Distributed B-format Recordings," *Acta Acustica*, vol. 5, paper 20 (2021 Apr.). https://doi.org/10.1051/aacus/2021012.

[9] E. Patricio, A. Ruminski, A. Kuklasinski, L. Januszkiewicz, and T. Zernicki, "Toward Six Degrees of Freedom Audio Recording and Playback using Multiple Ambisonics Sound Fields," presented at the *146th Convention of the Audio Engineering Society* (2019 Mar.), paper 10141.

[10] S. Spors, R. Rabenstein, and J. Ahrens, "The Theory of Wave Field Synthesis Revisited," presented at the *124th Convention of the Audio Engineering Society* (2008 May), paper 7358.

[11] S. Kaneko, T. Suenaga, H. Akiyama, Y. Miyake, S. Tominaga, F. Shirakihara, and H. Okumura, "Development of a 64-Channel Spherical Microphone Array and a 122-Channel Loudspeaker Array System for 3D Sound Field Capturing and Reproduction Technology Research," presented at the *144th Convention of the Audio Engineering Society* (2018 May), paper 10021.

[12] E. Stein and M. M. Goodwin, "Ambisonics Depth Extensions for Six Degrees of Freedom," in *Proceedings of the AES International Conference on Headphone Technology* (2019 Aug.), paper 23.

[13] J. G. Tylka and E. Y. Choueiri, "Performance of Linear Extrapolation Methods for Virtual Sound Field Navigation," *J. Audio Eng. Soc.*, vol. 68, no. 3, pp. 138–156 (2020 Mar.). https://doi.org/10.17743/jaes.2019.0054.

[14] F. Winter, F. Schultz, and S. Spors, "Localization Properties of Data-based Binaural Synthesis including Translatory Head-Movements," presented at the *Forum Acusticum*(Kraków, Poland) (2014 Sep.).

[15] N. Hahn and S. Spors, "Localization Properties of Data-based Binaural Synthesis including Translatory Head-Movements," presented at the *German Annual Conference on Acoustics (DAGA)*, pp. 1122–1125 (Nürnberg, Germany) (2015 Mar.).

[16] T. Pihlajamaki and V. Pulkki, "Synthesis of Complex Sound Scenes with Transformation of Recorded Spatial Sound in Virtual Reality," *J. Audio Eng. Soc.*, vol. 63, nos. 7–8, pp. 542–551 (2015 Aug.). https://doi.org/10.17743/jaes.2015.0059.

[17] J. G. Tylka and E. Y. Choueiri, "Soundfield Navigation using an Array of Higher-Order Ambisonics Microphones," in *Proceedings of the 2016 AES International Conference on Audio for Virtual and Augmented Reality (AVAR)* (2016 Sep.), paper 4-2.

[18] K. Wakayama, J. Trevino, H. Takada, S. Sakamoto, and Y. Suzuki, "Extended Sound Field Recording using Position Information of Directional Sound Sources," in *Proceedings of IEEE Workshop on Applications of*

Signal Processing to Audio and Acoustics (WASPAA), pp. 185–189 (New Paltz, New York) (2017 Oct.). https://doi.org/10.1109/WASPAA.2017.8170020.

[19] Y. Wang and K. Chen, "Translations of Spherical Harmonics Expansion Coefficients for a Sound Field using Plane Wave Expansions," *J. Acoust. Soc. Am.*, vol. 143, no. 6, pp. 3474–3478 (2018 Jun.). https://doi.org/10.1121/1.5041742.

[20] A. Plinge, S. J. Schlecht, O. Thiergart, T. Robotham, O. Rummukainen, and E. A. Habets, "Six-Degrees-of-Freedom Binaural Audio Reproduction of First-Order Ambisonics with Distance Information," in *Proceedings of the AES International Conference on Audio for Virtual and Augmented Reality* (2018 Aug.), paper P6-2.

[21] M. Kentgens, A. Behler, and P. Jax, "Translation of a Higher Order Ambisonics Sound Scene Based on Parametric Decomposition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 151–155 (Barcelona, Spain) (2020 May).

[22] M. Kentgens and P. Jax, "Comparison of Methods for Plausible Sound Field Translation," in *Proceedings of DAGA—47th Annual Conference on Acoustics* (Vienna, Austria) (2021 Aug.).

[23] L. McCormack, A. Politis, and V. Pulkki, "Parametric Spatial Audio Effects Based on the Multi-Directional Decomposition of Ambisonic Sound Scenes," presented at the *Proceedings of the 24th International Conference on Digital Audio Effects (DAFx20in21)*, pp. 214–221 (2021 Sep.).

[24] L. Birnie, T. Abhayapala, V. Tourbabin, and P. Samarasinghe, "Mixed Source Sound Field Translation for Virtual Binaural Application With Perceptual Validation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1188–1203 (2021 Feb.). https://doi.org/10.1109/TASLP.2021.3061939.

[25] E. Gallo, N. Tsingos, and G. Lemaitre, "3D-Audio Matting, Postediting, and Rerendering from Field Recordings," *EURASIP J. Adv Signal Process.*, vol. 2007, paper 047970 (2007 Dec.). https://doi.org/10.1155/2007/47970.

[26] E. Gallo and N. Tsingos, "Extracting and Re-Rendering Structured Auditory Scenes from Field Recordings," in *Proceedings of the AES 30th International Conference on Intelligent Audio Environments* (2007 Mar.), paper 19.

[27] K. Niwa, T. Nishino, and K. Takeda, "Encoding Large Array Signals into a 3D Sound Field Representation for Selective Listening Point Audio Based on Blind Source Separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 181–184 (Las Vegas, Nevada) (2008 Mar.). https://doi.org/10.1109/ICASSP.2008.4517576.

[28] N. Mariette and B. Katz, "SoundDelta–Largescale, Multi-User Audio Augmented Reality," in *Proceedings of the EAA Symposium on Auralization*, pp. 37–42 (Espoo, Finland) (2009 Jun.).

[29] C. Verron, P.-A. Gauthier, J. Langlois, and C. Guastavino, "Spectral and Spatial Multichannel Analysis/Synthesis of Interior Aircraft Sounds," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 7, pp. 1317–1329 (2013 Feb.). https://doi.org/10.1109/TASL.2013.2248712.

[30] O. Thiergart, G. Del Galdo, M. Taseska, and E. A. Habets, "Geometry-Based Spatial Sound Acquisition using Distributed Microphone Arrays," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 12, pp. 2583–2594 (2013 Aug). https://doi.org/10.1109/TASL.2013.2280210.

[31] X. Zheng, *Soundfield Navigation: Separation, Compression and Transmission*, Ph.D. thesis, School of Electrical, Computer, and Telecommunications Engineering, University of Wollongong, Wollongong, Australia (2013 Nov.).

[32] P. Samarasinghe, T. Abhayapala, and M. Poletti, "Wavefield Analysis Over Large Areas Using Distributed Higher Order Microphones," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 3, pp. 647–658 (2014 Jan.). https://doi.org/10.1109/TASLP.2014.2300341.

[33] E. Fernandez-Grande, "Sound Field Reconstruction using a Spherical Microphone Array," *J. Acoust. Soc. Am.*, vol. 139, no. 3, pp. 1168–1178 (2016 Mar.). https://doi.org/10.1121/1.4943545.

[34] S. Emura, "Sound Field Estimation using Two Spherical Microphone Arrays," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 101–105 (2017 Mar.). https://doi.org/10.1109/ICASSP.2017.7952126.

[35] E. Bates, H. O'Dwyer, K.-P. Flachsbarth, and F. M. Boland, "A Recording Technique for 6 Degrees of Freedom VR," presented at the *144th Convention of the Audio Engineering Society* (2018 May), paper 10022.

[36] D. R. Méndez, C. Armstrong, J. Stubbs, M. Stiles, and G. Kearney, "Practical Recording Techniques for Music Production with Six-Degrees of Freedom Virtual Reality," presented at the *145th Convention of the Audio Engineering Society* (2018 Oct.), paper 464.

[37] N. Ueno, S. Koyama, and H. Saruwatari, "Sound Field Recording Using Distributed Microphones Based on Harmonic Analysis of Infinite Order," *IEEE Signal Process Lett*, vol. 25, no. 1, pp. 135–139 (2018 Nov.). https://doi.org/10.1109/LSP.2017.2775242.

[38] M. Nakanishi, N. Ueno, S. Koyama, and H. Saruwatari, "Two-Dimensional Sound Field Recording With Multiple Circular Microphone Arrays Considering Multiple Scattering," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 368–372 (New Paltz, New York) (2019 Oct.). https://doi.org/10.1109/WASPAA.2019.8937208.

[39] C. Schörkhuber, F. Zotter, and R. Höldrich, "Triplet-Based Variable-Perspective (6DoF) Audio Rendering from Simultaneous Surround Recordings Taken at Multiple Perspectives," in *Proceedings of the 46th Annual Conference on Acoustics (DAGA)* (Hannover, Germany) (2020 Mar.).

[40] F. Zotter, M. Frank, C. Schörkhuber, and R. Höldrich, "Signal-Independent Approach to Variable-Perspective (6DoF) Audio Rendering from Simultaneous Surround Recordings taken at Multiple Perspectives," in *Proceedings of the 46th Annual Conference on Acoustics (DAGA)* (Hannover, Germany) (2020 Mar.).

[41] M. Pezzoli, F. Borra, F. Antonacci, S. Tubaro, and A. Sarti, "A Parametric Approach to Virtual Miking for Sources of Arbitrary Directivity," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2333–2348 (2020 Jul.). https://doi.org/10.1109/TASLP.2020.3012058.

[42] T. Ciotucha, A. Ruminski, T. Zernicki, and B. Mróz, "Evaluation of Six Degrees of Freedom 3D Audio Orchestra Recording and Playback using Multi-point Ambisonics interpolation," presented at the *150th Convention of the Audio Engineering Society* (2021 May), paper 10459.

[43] N. Iijima, S. Koyama, and H. Saruwatari, "Binaural Rendering from Microphone Array Signals of Arbitrary Geometry," *J. Acoust. Soc. Am.*, vol. 150, no. 4, pp. 2479–2491 (2021 Oct.). https://doi.org/10.1121/10.0006538.

[44] S. Kaneko and R. Duraiswami, "Multiple Scattering Ambisonics: Three-Dimensional Sound Field Estimation using Interacting Spheres," *arXiv preprint arXiv:2106.07157* (2021).

[45] M. McCrea, L. McCormack, and V. Pulkki, "Sound Source Localization Using Sector-Based Analysis With Multiple Receivers," in *Proceedings of the 2nd Nordic Sound and Music Conference* (Copenhagen, Denmark) (2021 Nov.).

[46] J. G. Tylka and E. Choueiri, "Comparison of Techniques for Binaural Navigation of Higher-Order Ambisonic Soundfields," presented at the *139th Convention of the Audio Engineering Society* (2015 Oct.), paper 9421.

[47] J. G. Tylka and E. Y. Choueiri, "Domains of Practical Applicability for Parametric Interpolation Methods for Virtual Sound Field Navigation," *J. Audio Eng. Soc.*, vol. 67, no. 11, pp. 882–893 (2019 Nov.). https://doi.org/10.17743/jaes.2019.0038.

[48] J. G. Tylka and E. Y. Choueiri, "Fundamentals of a Parametric Method for Virtual Navigation within an Array of Ambisonics Microphones," *J. Audio Eng. Soc.*, vol. 68, no. 3, pp. 120–137 (2020 Mar.). https://doi.org/10.17743/jaes.2019.0055.

[49] A. Laborie, R. Bruno, and S. Montoya, "A New Comprehensive Approach of Surround Sound Recording," presented at the *114th Convention of the Audio Engineering Society* (2003 Mar.), paper 5717.

[50] V. Pulkki, A. Politis, M.-V. Laitinen, J. Vilkamo, and J. Ahonen, "First-Order Directional Audio Coding (DirAC)," in V. Pulkki, S. Delikaris-Manias, and A. Politis (Eds.), *Parametric Time-Frequency Domain Spatial Audio*, pp. 89–138 (John Wiley & Sons, Ltd., Hoboken, New Jersey, USA, 2017), 1st ed. https://doi.org/10.1002/9781119252634.ch5.

[51] A. Allen and B. W. Kleijn, "Ambisonic Soundfield Navigation using Directional Decomposition and Path Distance Estimation," in *Proceedings of the 4th International Conference on Spatial Audio (ICSA)*, pp. 117–122 (Graz, Austria) (2017 Sept.).

[52] A. Politis, M.-V. Laitinen, J. Ahonen, and V. Pulkki, "Parametric Spatial Audio Processing of Spaced Microphone Array Recordings for Multichannel Reproduction," *J. Audio Eng. Soc.*, vol. 63, no. 4, pp. 216–227 (2015 Apr.). https://doi.org/10.17743/jaes.2015.0015.

[53] K. Müller and F. Zotter, "Auralization Based on Multi-Perspective Ambisonic Room Impulse Responses," *Acta Acust.*, vol. 4, no. 6, paper 25 (2020 Nov.). https://doi.org/10.1051/aacus/2020024.

[54] K. Han and A. Nehorai, "Improved Source Number Detection and Direction Estimation With Nested Arrays and ULAs Using Jackknifing," *IEEE Trans. Signal Process.*, vol. 61, no. 23, pp. 6118–6128 (2013 Sep.). https://doi.org/10.1109/TSP.2013.2283462.

[55] N. Epain and C. T. Jin, "Spherical Harmonic Signal Covariance and Sound Field Diffuseness," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 10, pp. 1796–1807 (2016 Jun.). https://doi.org/10.1109/TASLP.2016.2585862.

[56] O. Nadiri and B. Rafaely, "Localization of Multiple Speakers under High Reverberation using a Spherical Microphone Array and the Direct-Path Dominance Test," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 10, pp. 1494–1505 (2014 Jul.). https://doi.org/10.1109/TASLP.2014.2337846.

[57] M. Cobos, A. Marti, and J. J. Lopez, "A Modified SRP-PHAT Functional for Robust Real-Time Sound Source Localization With Scalable Spatial Sampling," *IEEE Signal Process. Lett.*, vol. 18, no. 1, pp. 71–74 (2010 Nov.). https://doi.org/10.1109/LSP.2010.2091502.

[58] O. L. Frost, "An Algorithm for Linearly Constrained Adaptive Array Processing," *Proceedings of the IEEE*, vol. 60, no. 8, pp. 926–935 (1972 Aug.). https://doi.org/10.1109/PROC.1972.8817.

[59] R. Schmidt, "Multiple Emitter Location and Signal Parameter Estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280 (1986 Mar.). https://doi.org/10.1109/TAP.1986.1143830.

[60] M. Kaveh and A. Barabell, "The Statistical Performance of the MUSIC and the Minimum-Norm Algorithms in Resolving Plane Waves in Noise," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 2, pp. 331–341 (1986 Apr.). https://doi.org/10.1109/TASSP.1986.1164815.

[61] L. McCormack and S. Delikaris-Manias, "Parametric First-Order Ambisonic Decoding for Headphones Utilising the Cross-Pattern Coherence Algorithm," in *Proceedings of the EAA Spatial Audio Signal Processing Symposium*, pp. 173–178 (Paris, France) (2019 Sep.). https://doi.org/10.25836/sasp.2019.26.

[62] Y. Yamasaki and T. Itow, "Measurement of Spatial Information in Sound Fields by Closely Located Four Point Microphone Method," *J. Acoust. Soc. Japan (E)*, vol. 10, no. 2, pp. 101–110 (1989 Mar.).

[63] F. J. Fahy and V. Salmon, "Sound Intensity," *J. Acoust. Soc. Am.*, vol. 88, no. 4, pp. 2044–2045 (1990 Oct.). https://doi.org/10.1121/1.400195.

[64] C. Schörkhuber and R. Höldrich, "Linearly and Quadratically Constrained Least-Squares Decoder for Signal-Dependent Binaural Rendering of Ambisonic Signals," in *Proceedings of the AES International Conference on Immersive and Interactive Audio* (2019 Mar.), paper 22.

[65] L. McCormack, S. Delikaris-Manias, A. Politis, et al., "Applications of Spatially Localized Active-Intensity Vectors for Sound-Field Visualization," *J. Au-*

dio Eng. Soc., vol. 67, no. 11, pp. 840–854 (2019 Nov.). https://doi.org/10.17743/jaes.2019.0041.

[66] B. Jo and J.-W. Choi, "Parametric Direction-of-Arrival Estimation with Three Recurrence Relations of Spherical Harmonics," *J. Acoust. Soc. Am.*, vol. 145, no. 1, pp. 480–488 (2019 Jan.). https://doi.org/10.1121/1.5087698.

[67] A. Herzog and E. A. Habets, "On the Relation Between DOA-Vector Eigenbeam ESPRIT and Subspace Pseudointensity-vector," in *Proceedings of the 27th European Signal Processing Conference (EUSIPCO)*, pp. 1–5 (A Coruna, Spain) (2019 Sep.). https://doi.org/10.23919/EUSIPCO.2019.8902715.

[68] B. Jo, F. Zotter, and J.-W. Choi, "Extended Vector-Based EB-ESPRIT Method," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1692–1705 (2020 May). https://doi.org/10.1109/TASLP.2020.2996090.

[69] L. McCormack, A. Politis, S. Särkkä, and V. Pulkki, "Real-Time Tracking of Multiple Acoustical Sources Utilising Rao-Blackwellised Particle Filtering," in *Proceedings of the 29th European Signal Processing Conference (EUSIPCO)*, pp. 206–210 (Dublin, Ireland) (2021 Aug.). https://doi.org/10.23919/EUSIPCO54536.2021.9616095.

[70] Y. Bar-Shalom, X.-R. Li, T. Kirubarajan and, *Estimation with Applications to Tracking and Navigation* (John Wiley & Sons, Inc., Hoboken, New Jersey, 2001). https://doi.org/10.1002/0471221279.

[71] S. Särkkä, A. Vehtari, and J. Lampinen, "Rao-Blackwellized Particle Filter for Multiple Target Tracking," *Inform. Fusion*, vol. 8, no. 1, pp. 2–15 (2007 Jan.). https://doi.org/10.1016/j.inffus.2005.09.009.

[72] J. Hartikainen and S. Särkkä, "RBMCDAbox–MATLAB Toolbox of Rao-Blackwellized Data Association Particle Filters," Documentation of RBMCDA Toolbox for MATLAB V (2008 Mar.) https://www.researchgate.net/publication/228914523_RBMCDAbox-Matlab_Toolbox_of_Rao-Blackwellized_Data_Association_Particle_Filters.

[73] B. Rafaely, *Fundamentals of Spherical Array Processing*, vol. 8 (Springer, Cham, Switzerland, 2015), 2nd ed. https://doi.org/10.1007/978-3-319-99561-8.

[74] S. Delikaris-Manias and V. Pulkki, "Cross Pattern Coherence Algorithm for Spatial Filtering Applications Utilizing Microphone Arrays," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 11, pp. 2356–2367 (2013 Aug.). https://doi.org/10.1109/TASL.2013.2277928.

[75] R. O. Duda and W. L. Martens, "Range Dependence of the Response of a Spherical Head Model," *J. Acoust. Soc. Am.*, vol. 104, no. 5, pp. 3048–3058 (1998 Nov.).

[76] J. Daniel, "Spatial Sound Encoding Including Near Field Effect: Introducing Distance Coding Filters and a Viable, New Ambisonic Format," presented at the *23rd AES International Conference on Signal Processing in Audio Recording and Reproduction* (2003 May), paper 16.

[77] C. Schörkhuber, M. Zaunschirm, and R. Höldrich, "Binaural Rendering of Ambisonic Signals via Magnitude Least Squares," in *Proceedings of the German Annual Conference on Acoustics (DAGA)*, vol. 44, pp. 339–342 (Munich, Germany) (2018 Mar.).

[78] T. McKenzie, D. T. Murphy, and G. Kearney, "Interaural Level Difference Optimization of Binaural Ambisonic Rendering," *Appl. Sci.*, vol. 9, no. 6, paper 1226 (2019 Mar.). https://doi.org/10.3390/app9061226.

[79] J. Vilkamo and T. Bäckström, "Time–Frequency Processing: Methods and Tools," in V. Pulkki, S. Delikaris-Manias, and A. Politis (Eds.), *Parametric Time-Frequency Domain Spatial Audio*, p. 3 (John Wiley & Sons, Hoboken, New Jersey, 2017), 1st ed. https://doi.org/10.1002/9781119252634.ch1.

[80] M. Bouéri and C. Kyriakakis, "Audio Signal Decorrelation based on a Critical Band Approach," presented at the *117th Convention of the Audio Engineering Society* (2004 Oct.), paper 6291.

[81] E. Kermit-Canfield and J. Abel, "Signal Decorrelation using Perceptually Informed Allpass Filters," in *Proceedings of the 19th International Conference on Digital Audio Effects (DaFX)*, pp. 225–31 (Brno, Czech Republic) (2016 Sep.).

[82] J. Herre, H. Purnhagen, J. Breebaart, et al., "The Reference Model Architecture for MPEG Spatial Audio Coding," presented at the *118th Convention of the Audio Engineering Society* (2005 May), paper 6447.

[83] T. McKenzie, L. McCormack, and C. Hold, "Dataset of Spatial Room Impulse Responses in a Variable Acoustics Room for Six Degrees-of-Freedom Rendering and Analysis," *arXiv preprint arXiv:2111.11882* (2021). https://doi.org/10.48550/arXiv.2111.11882.

[84] J. L. Hintze and R. D. Nelson, "Violin Plots: A Box Plot-Density Trace Synergism," *Am. Statistician*, vol. 52, no. 2, pp. 181–184 (1997 Feb.). https://doi.org/10.1080/00031305.1998.10480559.

[85] T. McKenzie, S. J. Schlecht, and V. Pulkki, "Auralisation of the Transition between Coupled Rooms," in *Proceedings of Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, pp. 1–9 (Bologna, Italy) (2021 Sep.). https://doi.org/10.1109/I3DA48870.2021.9610955.

[86] L. McCormack, A. Politis, and V. Pulkki, "Rendering of Source Spread for Arbitrary Playback Setups Based on Spatial Covariance Matching," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (New Paltz, New York) (2021 Oct.). https://doi.org/10.1109/WASPAA52581.2021.9632724.

THE AUTHORS

Leo McCormack

Archontis Politis

Thomas McKenzie

Christoph Hold

Ville Pulkki

Leo McCormack is a doctoral candidate for the Department of Signal Processing and Acoustics at Aalto University, Finland, researching parametric spatial audio technologies. He received his M.Sc. degree in Computer Communications and Information Sciences, majoring in Acoustics and Audio Technology, at Aalto University, Finland, and his B.Sc. in Music Technology and Audio Systems at the University of Huddersfield, UK. He was also a student intern at Fraunhofer IIS, Erlangen, Germany, in 2013–2014. His research interests include multichannel and microphone array signal processing for sound-field reproduction and sound source localization. He is also a strong advocate for open-source software and contributes to a number of open-source projects related to spatial audio.

•

Archontis Politis is a postdoctoral researcher at Tampere University, Finland. He obtained his M.Sc. degree in Sound & Vibration studies from the Institute of Sound and Vibration Research (ISVR), University of Southampton, UK, in 2008. In 2015, he was a visiting researcher at the University of Maryland Institute for Advanced Computer Studies, MA, USA, and in the same year, he completed a research internship at Microsoft Research, Redmond, WA, USA. In 2016, he obtained a Doctor of Science degree on spatial audio processing from Aalto University, Finland. He has served as editor of a book on Parametric Spatial Audio Processing, organizer in the DCASE scientific challenge, and has chaired various special sessions in international conferences. His research interests include spatial audio technologies, virtual acoustics, array signal processing, and acoustic scene analysis.

•

Thomas McKenzie is a postdoctoral researcher in the Department of Signal Processing and Acoustics at Aalto University where he studies room acoustics and six degrees-of-freedom spatial audio. He completed a B.Sc. in Music, Multimedia and Electronics at the University of Leeds, UK, in 2013, before completing a M.Sc. in Postproduction with Sound Design and a Ph.D. in Music Technology at the University of York, UK, in 2015 and 2020, respectively. His research interests include spatial audio and psychoacoustics.

•

Christoph Hold is a doctoral candidate in the Department of Signal Processing and Acoustics at Aalto University, Finland, focusing on spatial audio processing. He received an M.Sc. in audio communication technology in 2019 and a B.Sc. in electrical engineering from the Technische Universität Berlin, where he specialized in signal processing and virtual acoustics. From 2015 to 2017, he was a research assistant at TU Berlin, followed by two research internships (2017 and 2018) at Microsoft Research in Redmond, WA, USA. He is interested in high-quality audio and its perception. For the Audio Engineering Society, he was the Chair of the Berlin Student Section and part of the 142nd AES Convention committee.

•

Ville Pulkki is a professor in the Department of Signal Processing and Acoustics at Aalto University, Helsinki, Finland. He has been working in the field of spatial audio for over 20 years. He developed the vector-base amplitude panning (VBAP) method in his Ph.D. (2001) and directional audio coding after his Ph.D. with his research group. He also has contributions in perception of spatial sound, laser-based measurement of room responses, and binaural auditory models. He has received the Samuel L. Warner Memorial Medal Award from SMPTE and the AES Silver Medal Award. He enjoys being with his family, building his summer house, and performing in musical ensembles.