Acta Acustica 2021, **5**, 20

https://doi.org/10.1051/aacus/2021012

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

SCIENTIFIC ARTICLE OPEN ACCESS

Particle-filter tracking of sounds for frequency-independent 3D audio rendering from distributed B-format recordings

Matthias Blochberger* and Franz Zotter

Institute of Electronic Music and Acoustics, University of Music and Performing Arts Graz, Inffeldgasse 10/III, 8010 Graz, Austria

Received 2 October 2020, Accepted 16 March 2021

Abstract – Six-Degree-of-Freedom (6DoF) audio rendering interactively synthesizes spatial audio signals for a variable listener perspective based on surround recordings taken at multiple perspectives distributed across the listening area in the acoustic scene. Methods that rely on recording-implicit directional information and interpolate the listener perspective without the attempt of localizing and extracting sounds often yield high audio quality, but are limited in spatial definition. Methods that perform sound localization, extraction, and rendering typically operate in the time-frequency domain and risk introducing artifacts such as musical noise. We propose to take advantage of the rich spatial information recorded in the broadband time-domain signals of the multitude of distributed first-order (B-format) recording perspectives. Broadband time-variant signal extraction retrieving direct signals and leaving residuals to approximate diffuse and spacious sounds is less of a quality risk, and likewise is the broadband re-encoding to enhance spatial definition of both signal types. To detect and track direct sound objects in this process, we combine the directional data recorded at the single perspectives into a volumetric multi-perspective activity map for particle-filter tracking. Our technical and perceptual evaluation confirms that this kind of processing enhances the otherwise limited spatial definition of direct-sound objects of other broadband but signal-independent virtual loudspeaker object (VLO) or Vector-Based Intensity Panning (VBIP) interpolation approaches.

Keywords: 6DoF rendering, Variable-perspective rendering, Multi-perspective audio

1 Introduction

The interactive rendering of recorded auditory scenes as virtual listening environments requires an approach to allow six Degrees of Freedom (6DoF) of movement for a variable listener perspective. The variable-perspective rendering of auditory scenes requires interpolation between static recording perspective positions. In existing research, this concept is often referred to as scene navigation or also scene walk-through. This contribution mainly refers to first-order tetrahedral microphone arrays as means for recording surround audio for high fidelity applications.

While volumetrically navigable 6DoF recording and rendering are theoretically feasible, practical distributions of multiple static 3D audio recordings typically consider capturing perspective changes along the horizontal dimensions to enable walkable rendering of the auditory scene.

Perspective extrapolation of a single perspective for a shifted listening position has been considered in the SpaMoS (spatially modified synthesis) method by Pihlajamäki and Pulkki [1, 2] that estimates time-frequency-domain source positions by projecting directional signal detections of DirAC (directional audio coding [3, 4]) onto a pre-defined convex hull (e.g. the room walls). This method is assumed to be accurate in spatial reproduction when the parallactic shift with regard to the original perspective recording stays small. Similarly, Plinge et al. [5] utilize DirAC in combination with known distance information to rotate and attenuate sources in a single-perspective recording to extrapolate its perspective. This approach would be expandable by multiple directional signals obtained via HARPEX by Barrett and Berge [6] for first-order signals, as Stein and Goodwin suggest in [7]. Higher-order Ambisonics signals are used in a work by Allen and Kleijn [8] that employs a matching-pursuit algorithm for multi-directional signal decomposition, takes into account estimated source distances, and labels reflections and direct sounds. Kentgens et al. [9] apply alternative multi-directional decomposition, e.g., the subspace methods SORTE/MUSIC for extraction and shifted re-encoding of direct components, complemented by a noise and diffuseness subspace that preserves ambient sounds. Birnie et al. [10] introduce a sound field translation method for 6DoF binaural rendering based on sparse plane-wave expansion for near and far sources arranged on two rings around the higher-order recording perspective. Altogether, the single-perspective extrapolation approaches require either parametric time-frequency processing or higher order microphone arrays to achieve

*Corresponding author: matthias.blochberger@posteo.at

directional definition, and their extrapolation range depends on how successful distance information is guessed or estimated. Alternatively, Bates and O'Dwyer [11], Lee et al. [12] employ a more classical, spaced array augmented by controllable-directivity microphones to simulate an extrapolated listening perspective.

Multiple perspectives contain additional information needed for explicit acoustic source localization that enlarges the supported range of shifted listening perspectives with high spatial definition. Brutti et al. [13, 14], Hack [15] or Del Galdo et al. [16, 17] introduce object localization methods using maps of the acoustic activity to localize or triangulate the sources within a scene. In [13–15], sequential peak picking algorithms are proposed to avoid erroneous detections in correlation- and intensity-based triangulation respectively. The virtual microphone method [16, 17] utilizes the detected location every short-time frequency bin to assemble a virtual microphone signal of an arbitrary parametrically rendered directivity pattern, what in principle would allow parametric 6DoF rendering. Zheng [18] combines sound object detection with time-frequency-domain signal separation to extract direct signals for sound field navigation. Particle filters to track sound sources detected in reverberant environments were described by Ward et al. [19], and Fallon et al. [20] introduce a particle filter-based acoustic source tracking algorithm for a time-varying number of sources. Probabilistic detection was combined with particle filters for directional tracking of sounds was introduced by Valin et al. [21, 22] for robotic applications, which was later adapted by Kitić and Guérin [23] to a first-order Ambisonics application.

Multi-perspective recordings also permit perspective interpolation methods with information about source locations staying implicit. Tylka and Choueiri [24–26] proposed interpolatory spherical-harmonic re-expansion at low frequencies for multi-perspective first- or higher-order Ambisonic recordings. Similar to the simplified treatment of high-frequencies in Tylka et al., there is also a number of simplistic broadband signal interpolation methods working in the time domain, only. Such methods avoid any risk of introducing musical noise artifacts, which can happen in time-frequency-domain processing. Mariette et al. [27] mix the first-order Ambisonics signals of the three nearest recording positions proportional to their proximity, similar to the proposal of Schörkhuber et al. [28] that introduces additional amplitude-dependent gains. Patricio et al. [29] propose a distance-based linear interpolation between higher-order Ambisonic recordings in the time domain, in which the higher-order content of distant microphones is faded out, while proximal recording perspectives remain unaltered. Multiple perspective recordings are mapped to multiple surround playback rings of virtual loudspeaker objects (VLOs) by Grosche et al. [30] and Zotter et al. [31], of which the direction and amplitude (involving a distance and directivity function) vary with the listener position and are employed in higher-order re-encoding of the VLO signals. A good introduction to scene recording and sound field interpolation especially in multimedia VR applications is given by Rivas-Mendez et al. [32].

For simplicity, simulations and experiments in this contribution deal with an equidistant grid of recording perspectives volumetrically distributed within a homogeneous auditory scene, however the approach introduced is more general. We introduce a multi-perspective interpolation method that merges and extends detection/tracking and broad-band signal processing concepts found in literature. A broadband signal extraction and rendering method is utilized for artifact-free signal processing in combination with automatic signal detection and position estimation for higher spatial accuracy. The estimated position of any detected object is used to steer broadband beamformers at the nearest recording positions to capture the object's direct sound. Weighted and delay-compensated combinations of the extracted signals yield approximated direct signals, while residual signals with direct-sound directions suppressed aim to reintroduce enveloping components of the diffuse sound field. Signal extraction and encoding procedures are described in Section 2 and scene analysis procedures in Section 3. Detection accuracy is technically investigated in Section 4 under varying SNR conditions. To assess the performance and achievable improvement of the proposed algorithm applied to a simple acoustic scene recording with static objects, a two-part listening experiment compares the rendering method with two existing broadband 6DoF rendering methods in Section 5, for a static and a moving listener.

2 Frequency-independent 6DoF Rendering

Given the listener position, the microphone array positions, and assuming to know the sound source positions, we can compute the signals customized to the acoustic perspective of a single virtual listener, delivered as a stream of higher-order Ambisonics [33] signals. This delivery format features the benefits of modular decoders facilitating playback on headphones or loudspeaker layouts and the Ambisonic sound field rotation operator. In case there are multiple virtual listeners, the listening-position-dependent processing steps are carried out interactively and separately for each listener, excluding the sound-scene analysis steps that are pre-computed offline.

2.1 Signal encoding and decoding

The multi-channel Ambisonics signal $\chi(t)$ of the listener perspective of the order N is computed by multiplication of a single-channel signal with the weights of an encoder $\boldsymbol{y}_{\mathrm{N}}(\theta)$. Such an encoder consists of the N3D-normalized spherical harmonics $\boldsymbol{y}_N(\boldsymbol{\theta})^{\mathrm{T}} = [Y_0^0(\boldsymbol{\theta}), Y_1^{-1}(\boldsymbol{\theta}), \ldots, Y_1^1(\boldsymbol{\theta}), \ldots, Y_N^N(\boldsymbol{\theta})]$ evaluated at the unit-length direction vector $\theta = [\cos\varphi \sin\vartheta, \sin\varphi \sin\vartheta, \cos\vartheta]$. The theoretical background of spherical harmonics and the concept of Ambisonics can be found in literature, e.g. [33], and practical implementation of encoders and decoders alike is easily accomplished with libraries such as introduced by [34].[1] We will encode the

[1] MATLAB implementation *Spherical-Harmonic-Transform* by Archontis Politis available at https://github.com/polarch/Spherical-Harmonic-Transform.

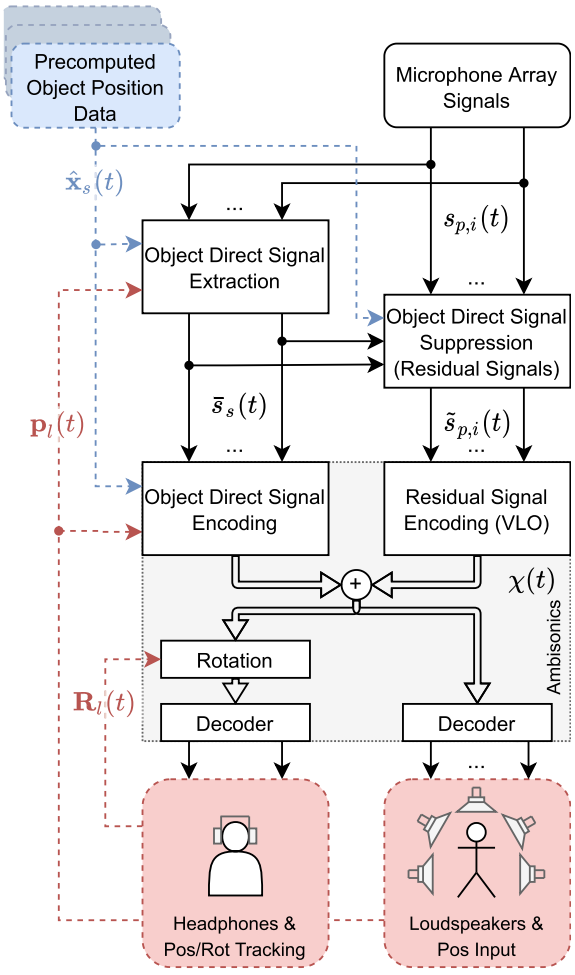


Figure 1. Diagram of the rendering algorithm. Position data (cf. Sect. 3) of sound objects is used to render the Ambisonics listener perspective using real time tracking of the listener position and head rotation.

listening-position-dependent object direct signals (cf. Sect. 2.2) and the residual signals (cf. Sect. 2.3), depending on the relative direction vector θ to the listener. For S signals in total, the signals $s_i(t)$, their gains g_i, and the instantaneous direction $\boldsymbol{\theta}_i = \boldsymbol{\theta}_i(t)$ of each signal, encoding is defined as

$$\boldsymbol{\chi}(t) = \sum_{i=1}^{S} \boldsymbol{y}_N(\boldsymbol{\theta}_i)\, s_i(t)\, g_i. \tag{1}$$

Rotation of the Ambisonics perspective is necessary with headphone playback (Fig. 1). It enables taking into account the listener's head rotation to simulate a static acoustic scene outside the headphones. This is done by applying a $(N+1)^2 \times (N+1)^2$ rotation matrix $R(\varphi, \vartheta, \gamma)$ (cf. [35, 33]) to the Ambisonics signal

$$\tilde{\boldsymbol{\chi}}(t) = \boldsymbol{\chi}(t) R(\varphi, \vartheta, \gamma). \tag{2}$$

To render the headphone signals from the variable-rotation Ambisonics signal $\tilde{\boldsymbol{\chi}}$, the *MagLS* binaural decoder is used, as described in [36]. With the exception of this binaural decoder, no frequency-dependent signal processing is introduced by the signal processing proposed.

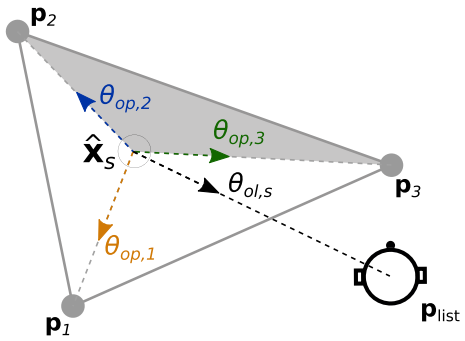


Figure 2. Three perspective extraction.

The upcoming sections introduce the methods to extract object direct signals and residual signals from the multi-perspective recordings, and they define gain and direction values for the encoding step in (1).

2.2 Object direct signal extraction

The object direct signals are approximations of the signal emitted from audible objects in the scene arriving at the virtual listener. Their positions are assumed to be known here; Section 3 thereafter introduces position estimation.

Figure 2 shows the positions of three microphone arrays, an object position, and the listener position, which are all required to approximate the direct signal of the object as a weighted sum. In general, for each known or estimated sound object with position inside the scene, a simplex of surrounding microphone arrays is selected to define the closest set of surrounding recordings taken. The signals at its vertices are likely to capture the cleanest instances of the object's direct sound. Initial weights for the signals at these vertices are obtained from the area coordinates, the barycentric coordinates of a point in the triangle (Fig. 2) as the typical simplex when recording positions are distributed horizontally, cf. (7); or within a tetrahedron if recordings were distributed volumetrically. But first, the microphone arrays at any of these position are a spherical constellation of transducers, and $\boldsymbol{\theta}_{\mathrm{m},j}$ denotes the transducer direction vectors. In our case, the index range is $j = 1 \ldots 4$, as the arrays considered are tetrahedral Oktava MK 4012 [37]. The tetrahedral cardioid microphone signals of each recording perspective are encoded into N3D-normalized first-order Ambisonics by

$$\boldsymbol{\psi}(t) = \operatorname{diag}\Big\{\big[1,\ \sqrt{3},\ \sqrt{3},\ \sqrt{3}\big]\Big\} \sum_{j=1}^{4} \boldsymbol{y}_1\big(\boldsymbol{\theta}_{\mathrm{m},j}\big) s_{p,j}(t), \tag{3}$$

assuming the jth recorded transducer signals of the perspective p is denoted as $s_{p,j}(t)$. From this, a signal

$$s_{\mathrm{obj}}(t) \approx \boldsymbol{g}^{\mathrm{T}} \boldsymbol{\psi}(t), \tag{4}$$

is extracted by applying beamforming vector $\boldsymbol{g}$ that steers towards the object position, here denoted as $-\boldsymbol{\theta}_{\mathrm{op},i}$, which is illustrated in Figure 2. This beamforming vector is

modelled as on-axis-normalized maximum-directivity (in first order: hypercardioid) and can be computed as

$$\boldsymbol{g} = \frac{\boldsymbol{y}_1(-\boldsymbol{\theta}_{\mathrm{op},i})}{\boldsymbol{y}_1^{\mathrm{T}}(-\boldsymbol{\theta}_{\mathrm{op},i})\boldsymbol{y}_1(-\boldsymbol{\theta}_{\mathrm{op},i})} = \frac{\boldsymbol{y}_1(-\boldsymbol{\theta}_{\mathrm{op},i})}{||\boldsymbol{y}_1(-\boldsymbol{\theta}_{\mathrm{op},i})||^2}, \tag{5}$$

where $\boldsymbol{y}_1$ is the encoder of first order in the negative object-perspective direction $-\boldsymbol{\theta}_{\mathrm{op},i}$.

To combine the three beamforming-extracted signals of an object from the triplet $i = 1 \ldots 3$ of into a single direct signal, a combined gain is defined for the object s and perspective i

$$g_{s,i} = \sqrt{\frac{g_{\mathrm{tri},s,i}\, g_{\mathrm{dir},s,i}}{\sum_{j=1}^{3} g_{\mathrm{tri},s,j}\, g_{\mathrm{dir},s,j}}} \quad \text{with } i = 1 \ldots 3. \tag{6}$$

Herein, the gain $g_{\mathrm{dir},s,i}$ denotes the *areal* or *barycentric* coordinate weight, cf. [38]. Assuming a projected sound object position $\hat{\boldsymbol{x}}_{\mathrm{2D},s}$, it favors the closest perspective from the given triplet of projected positions $\boldsymbol{p}_{\mathrm{2D},1}$, $\boldsymbol{p}_{\mathrm{2D},2}$, $\boldsymbol{p}_{\mathrm{2D},3}$ and yields

$$\begin{bmatrix} g_{\mathrm{tri},s,2} \\ g_{\mathrm{tri},s,3} \end{bmatrix} = \mathbf{C}^{-1}\left(\hat{\boldsymbol{x}}_{\mathrm{2D},s} - \boldsymbol{p}_{\mathrm{2D},1}\right) \tag{7}$$

with

$$\mathbf{C} = [\boldsymbol{p}_{\mathrm{2D},2} - \boldsymbol{p}_{\mathrm{2D},1}\, \boldsymbol{p}_{\mathrm{2D},3} - \boldsymbol{p}_{\mathrm{2D},1}]. \tag{8}$$

The remaining value is computed by

$$g_{\mathrm{tri},s,1} = 1 - (g_{\mathrm{tri},s,2} + g_{\mathrm{tri},s,3}). \tag{9}$$

The factor $g_{\mathrm{dir},s,i}$ quantifies the alignment of the object-perspective directions $\boldsymbol{\theta}_{\mathrm{op},i}$ with the object-listener direction $\boldsymbol{\theta}_{\mathrm{ol},s}$. Its task is to favor perspectives that are directionally aligned with the direction of radiation from the source to the listener, and it hereby favors the most suitable surrounding perspectives in terms of source directivity. It is formalized as cardioid

$$g_{\mathrm{dir},s,i} = \left(\frac{1 + \boldsymbol{\theta}_{\mathrm{op},i}^{\mathrm{T}}\boldsymbol{\theta}_{\mathrm{ol},s}}{2}\right)^{\alpha} \tag{10}$$

of the order α. Again, the vectors $\boldsymbol{\theta}_{\mathrm{op},i}$ and $\boldsymbol{\theta}_{\mathrm{ol},s}$ are unit vectors and illustrated by Figure 2.

The signal encoding as introduced with (1) requires direction vector, signal and gain. The object-listener direction $\boldsymbol{\theta}_{\mathrm{ol},s}$ is employed to compute the encoder (Sect. 2.1) for the approximated object direct signal, which is the combination of the areal coordinate gain-weighted (6) and delay-compensated extracted signals (4)

$$\bar{s}_s(t) = \sum_{i=1}^{3} g_{s,i}\, s_{\mathrm{obj},i}(t - \Delta t_{s,i}). \tag{11}$$

Here, delay compensation is done based on the speed of sound c and distance differences $\Delta t_{s,u} = c^{-1}\,(d_{\mathrm{ol},s} - d_{\mathrm{op},s,i})$, where the distances $d_{\mathrm{ol},s}$ and $d_{\mathrm{op},s,i}$ denote the object-listener and object-perspective distances for the object s and perspective $i = 1 \ldots 3$ in the triplet, respectively, see Figure 2. To model a realistic distance-dependent amplitude attenuation of the signals within the Ambisonic listener perspective the areal coordinate gains are first multiplied by the object-perspective to object-listener distance ratio. Then the combination

$$g_s = \min\left\{4, \sum_{i=1}^{3} g_{s,i}\, \frac{d_{\mathrm{op},s,i}}{d_{\mathrm{ol},s}}\right\} \tag{12}$$

is the gain that is employed in (1) and depends on the distance between listener and source. It is limited to a maximum of 4 (+12 dB) to make avoid excessive boosts whenever $d_{\mathrm{ol},s}$ becomes small.

2.3 Object direct signal suppression (residual signals)

In the optimal case, the approximated object direct signals (11) exclude all room information such as early reflections and late reverberation. They provide a clean signal for accurate directional perception, however do not convey a realistic room impression to the listener. To this end, the residual signals are introduced. A similar concept of direct signal suppression, despite in the higher-order Ambisonics domain, was employed in [9] to extract ambient components.

Here, the concept of a residual signal is implemented in terms of the virtual loudspeaker object (VLO) approach [30, 31] that is illustrated in Figure 3a. Each perspective $\boldsymbol{p}_p$ holds a number of microphones of the directions $\boldsymbol{\theta}_{\mathrm{m},j}$. Each direction and microphone signal is represented by a VLO that is positioned at a finite distance R from the perspective, $\boldsymbol{p}_p + R\boldsymbol{\theta}_{\mathrm{m},j}$. The normalized direction vector from the listener position to the virtual loudspeaker objects as well as the corresponding distances are computed as

$$\phi_{p,i} = \frac{(\boldsymbol{p}_p + R\boldsymbol{\theta}_{\mathrm{m},j}) - \boldsymbol{p}_{\mathrm{list}}}{r_{p,i}} \tag{13}$$

and

$$r_{p,i} = \left|\left|(\boldsymbol{p}_p + R\boldsymbol{\theta}_{\mathrm{m},j}) - \boldsymbol{p}_{\mathrm{list}}\right|\right|, \tag{14}$$

respectively. As described in [30, 31], the VLO gains should depend on the distance

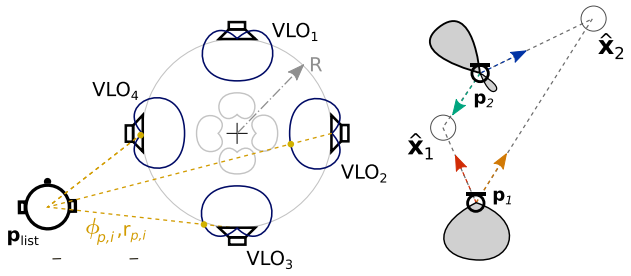
$$g_{\mathrm{dis},p,i}(r, R) = \begin{cases} \frac{R}{r_{p,i}}, & \text{for } r_{p,i} > R \\ \frac{r_{p,i}}{R}, & \text{for } r_{p,i} \le R \end{cases} \tag{15}$$

and in direction parametrically (from unity to cardioid)

$$g_{\mathrm{dir},p,i}(r) = \left(1 - \frac{\alpha}{2}\right) + \frac{\alpha}{2}\boldsymbol{\theta}_{\mathrm{m},j}^{\mathrm{T}}\phi_{p,i}, \tag{16}$$

$$\alpha = \frac{r_{p,i}}{r_{p,i} + R_{\mathrm{dir}}}. \tag{17}$$

Here, the gain (15) attenuates distant VLOs by $\frac{1}{r}$. Too close ones are attenuated by r to maintain a robust result and



(a) VLO (b) Direct signal de-emphasis gain pattern.

Figure 3. The VLO method [30, 31] is visualized in (a) while (b) shows the de-emphasis term introduced in eqaution (18).

erroneous localization. The gain (16) ensures that listeners do not hear sound far behind a VLO, which is always on-axis oriented towards $\boldsymbol{p}_\mathrm{p}$, but not abruptly so when walking through the VLO position. And the direction $\phi_{p,i}$ in (13) represents the parallactic displacement at a shifted listening position. The VLO approach achieves an enveloping and spatially plausible reproduction when used with multi-perspective microphone arrays distributed in the recorded scene. There is potential for improvement in the spatial definition of its direct-sound imaging.

For this work, the VLO method is now modified to serve as a residual-signal renderer complementing the objects direct sound signals. Here, the method encodes the $4P$ residual signals to the listener perspective using (13) as directions in (1). It is however necessary to exclude the direct signals from the VLO signals so the spatial accuracy gained by object signal encoding (Sect. 2.2) is not diminished. To this end, de-emphasis is applied to each microphone-array perspective with the goal of suppressing direct signals that belong to identified sound objects. This is done by gains depending on the object-perspective directions and the object's direct signal amplitude applied to the signals $s_{p,i}(t)$ for all perspectives $p = 1 \ldots P$ and transducers $i = 1 \ldots 4$. Extending the gains (15) and (16) introduced in [30, 31] by a de-emphasis term $G_p(t, \boldsymbol{\theta}_{\mathrm{m},i})$ leads to the new residual VLO gains

$$g_{\mathrm{r},p,i} = G_p(t, \boldsymbol{\theta}_{\mathrm{m},i}) g_{\mathrm{dis},p,i}\, g_{\mathrm{dir},p,i}. \tag{18}$$

Figure 3b illustrates the concept of this de-emphasis term suppressing direction towards the objects, which is the product of S directional gain patterns. Each of them is a mixture of unity and cardioid directivity, oriented such that the cardioid suppresses the object signals. It is defined for the array transducer directions $\boldsymbol{\theta}_{\mathrm{m},i}$ as

$$G_p(t, \boldsymbol{\theta}_{\mathrm{m},i}) = \prod_{s=1}^{S} \tilde{G}_{p,s}(t, \boldsymbol{\theta}_{\mathrm{m},i}) \tag{19}$$

with the single-object de-emphasis pattern

$$\tilde{G}_{p,s}(t, \boldsymbol{\theta}_{\mathrm{m},i}) = \big[1 - a_{p,s}(t)\big] + a_{p,s} \left(\frac{1 + \boldsymbol{\theta}_{\mathrm{op},p}^{\mathrm{T}}(t) \boldsymbol{\theta}_{\mathrm{m},i}}{2} \right)^{\beta}, \tag{20}$$

for which the exponent β controls the width of the directional notch and $a_{p,s}(t) \in [0,1]$ permits to control the depth of the notch, or to release it for distant or quiet sound-object signals. For this purpose, $a_{p,s}(t)$ is defined depending on the object-perspective distance and moving RMS value of $\bar{s}_s(t)$ (11) as

$$a_{p,s}(t) = \frac{s_{\mathrm{RMS},s}(t)}{\mathrm{RMS}_s}\, g_{\mathrm{dis}}\big(d_{\mathrm{op},p}, R_{\mathrm{de}}\big), \tag{21}$$

with

$$s_{\mathrm{RMS},s}(t) = \sqrt{\frac{1}{\Delta t_{\mathrm{RMS}}} \int_{t-\Delta t_{\mathrm{RMS}}}^{t} \bar{s}_s(\tau)^2 \mathrm{d}\tau}, \tag{22}$$

and $g_{\mathrm{dis}}(d_{\mathrm{op},p}, R_{\mathrm{de}})$ from 15 using reference distance R_{de}. The value of the threshold RMS_s for each sound object signal is determined by pre-computation given the recordings, or it can be defined manually.

3 Estimation of object positions

Figure 4 provides an overview of the procedure to estimate the sound object positions necessary for the rendering algorithm as introduced in previous sections. Given the frequency-domain microphone array surround signals, the direction-of-arrivals (DOAs) of single-frequency components are estimated and combined into DOA maps. This is similar in concept and application as the DOA histograms in [15, 18] and explained in Section 3.1. Section 3.2 introduces the method to intersect the directional information, to compute combined values, described as the acoustic activity map. Together with the subsequent sequential peak picking algorithm (Sect. 3.3), these concepts are also discussed in [15, 13, 14]. After selection of the instantaneous set of peaks by a sequential algorithm, these are evaluated in terms of probabilistic measures for sound object emergence, continuous activity and false detections. Section 3.7 explains the computation of these measures. They inform the decision making regarding the instantiation of particle filters for peak tracking. We use these particle filters, a well known Monte-Carlo probabilistic estimation technique and application-specifically described in Section 3.5, to track the three-dimensional position of the acoustic activity peaks as a time-coherent trajectory, expanding the method introduced in [22, 23].

Algorithm 1: The computational steps done at each time instant m.

for $\boldsymbol{m = 1 \ldots M}$ **do**
- (1) Single-perspective DOA maps (Sect. 3.1)
- (2) Multi-perspective acoustic activity map (Sect. 3.2)
- (3) Peak detection (Sect. 3.3)
- (4) Particle filter prediction (Sect. 3.5)
- (5) Object detection (Sect. 3.7)
- (6) Particle filter initialization/deletion (Sect. 3.4)
- (7) Particle filter re-sampling (Sect. 3.5)

End

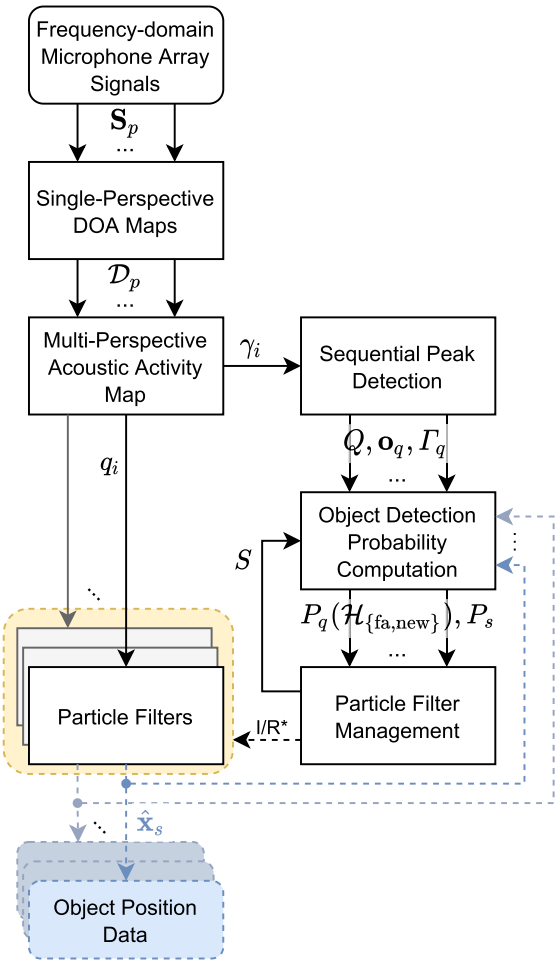


Figure 4. The procedure of object detection and object position estimation. $^*I/R$ signifies the decisions to initialize, continue or remove particle filter instances as introduced in Section 3.4.

3.1 Single-perspective DOA map

Direction-of-arrival estimation is applied frame-wise to the surround microphone array signals. The applied approach is the magnitude sensor response method, introduced in [39]. In [40] it was extended to the smoothed magnitude sensor response. It is applied in the frequency domain and done by frequency-wise computation of the covariance matrix $\mathbf{\Sigma}^{(k)}$ of the frequency-domain signal data $\boldsymbol{S}^{(k)}$ for bins $k = 1 \ldots K$ at the time instant m as an average over M time frames

$$\mathbf{\Sigma}^{(k)}(m) = \frac{1}{M} \sum_{m'=m-\frac{M}{2}}^{m+\frac{M}{2}-1} \boldsymbol{S}^{(k)}[m'] \cdot \boldsymbol{S}^{(k)}[m']^{\mathrm{H}}. \tag{23}$$

of the microphone array frequency bin magnitudes $\boldsymbol{S}^{(k)}$. A subsequent eigenvalue decomposition

$$\mathbf{\Sigma}^{(k)} = \boldsymbol{U}^{(k)} \mathbf{\Lambda}^{(k)} \boldsymbol{U}^{H(k)} \tag{24}$$

gives us the possibility to further decompose it into signal and noise subspace. This is done by selecting L eigen values $\lambda_l^{(k)}$ per bin k. As used in [40], this can be a fixed number L or be variable and computed with methods such as SORTE introduced in [41]. The case of a first-order surround signal allows the estimation of a maximum of two directions per frequency bin as it is the case with HARPEX [6]. The eigenvectors $\boldsymbol{u}_{*l}^{(k)}$, the columns of the left eigenvector matrix $\boldsymbol{U}^{(k)}$ corresponding to the selected L eigenvalues, span the signal subspace. These vectors give the estimation of the DOAs

$$\hat{\boldsymbol{\theta}}_l^{(k)} = \frac{\mathbf{V}\left|\mathbf{u}_{*l}^{(k)}\right|}{\left|\left|\mathbf{V}\left|\mathbf{u}_{*l}^{(k)}\right|\right|\right|} \quad \text{for } l = 1 \ldots L \tag{25}$$

for each frequency bin k.

In a similar application, [15] uses histograms accumulating DOA estimates. This concept is evolved into a non-discrete map in the spherical harmonics domain. The eigenvalues $\lambda_l^{(k)}$ and the corresponding DOAs $\hat{\boldsymbol{\theta}}_l^{(k)}$ of all frequency bins are aggregated into a broadband single-perspective DOA map $\mathcal{D}$ that is represented by real-valued spherical harmonics (cf. [33]) of the order N

$$\mathcal{D} = \sum_{k=1}^{K} \sum_{l=1}^{L} \boldsymbol{y}_N\left(\boldsymbol{\theta}_l^{(k)}\right) \mathcal{L}\left(k, \lambda_l^{(k)}\right). \tag{26}$$

The frequency- and eigenvalue-dependent function

$$\mathcal{L}(k, \lambda) = \begin{cases} k\sqrt{\lambda} & \text{if } \frac{k}{K} f_s > 200\,\text{Hz} \\ 0 & \text{else} \end{cases} \tag{27}$$

is used to limit the frequency range and compress large ranges of eigenvalues.

DOA maps according to (26) are continuous and have implicit smoothing that depends on the order, what permits interpolated evaluation at arbitrary directions. This is necessary for the intersection of the single-perspective DOA maps, sampled on a three-dimensional grid.

3.2 Multi-perspective acoustic activity map

The computation of three-dimensional data from the single-perspective DOA maps is based on work in [13–15] and further involves spherical harmonic representation/encoding as in Section 3.1 as well as decoding to interpolate for the subsequent computations. The continuous maps (26) obtained allow the computation of values for any direction, and in extension: position. Despite this multi-perspective fusion of DOA maps to sound object position implies somewhat omnidirectional sources, at least statistically, the rendering (Sect. 2) takes source directivity into account, as far as observable from the sparse recording perspectives. It does so in the position-dependent signal extraction (Sect. 2.2 and (10)) by favoring the perspectives best aligned with the radiation directions to the listener.

We introduce a set of positions $\boldsymbol{s}_i$ with $i = 1 \ldots G$, which can be regarded to be arbitrary for now and will be used in two different ways below. Then we define the unit directions

$$\boldsymbol{\theta}_{p,i} = \frac{\boldsymbol{s}_i - \boldsymbol{p}_p}{||\boldsymbol{s}_i - \boldsymbol{p}_p||}. \tag{28}$$

that sample the directions from each recording perspective position $\boldsymbol{p}_p$ to every position $\boldsymbol{s}_i$ in the set. The corresponding directions discretize the DOA map of any perspective by evaluating a spherical harmonic interpolation matrix

$$\tilde{\boldsymbol{Y}}_p = \begin{bmatrix} | & | & & | \\ \boldsymbol{y}_N(\boldsymbol{\theta}_{p,1}) & \boldsymbol{y}_N(\boldsymbol{\theta}_{p,2}) & \dots & \boldsymbol{y}_N(\boldsymbol{\theta}_{p,G}) \\ | & | & & | \end{bmatrix} \tag{29}$$

and hereby discretizing the single-perspective DOA map to display DOA activity for the positions $\boldsymbol{s}_\mathrm{i}$

$$\boldsymbol{w}_p = \tilde{\boldsymbol{Y}}_p \mathcal{D}_p. \tag{30}$$

These discrete DOA activities are subsequently weighted by a distance factor to give emphasis to perspectives close to the position $\boldsymbol{s}_i$,

$$f_p(i) = \exp\left(-\frac{||\boldsymbol{s}_i - \boldsymbol{p}_p||^2}{\delta_d}\right). \tag{31}$$

This function is applied element wise to $\boldsymbol{w}_p$, and $\delta_d = 2$ is the decay factor chosen here.

Next, the distance-weighted single-perspective values are combined by applying an l-norm

$$\gamma_i = \left[\sum_{p=1}^{P} \left(f_p(i)\boldsymbol{w}_p[i]\right)^l\right]^{\frac{1}{l}} \quad \text{for } i = 1 \dots G. \tag{32}$$

The term $\boldsymbol{w}_p[i]$ denotes the i-th element of the vector. This yields one value γ_i for each position $\boldsymbol{s}_i$ that will be subsequently called the *acoustic activity* for the positions in question. The acoustic activities $\{\gamma_i\}$ for an entire set of positions $\{\boldsymbol{s}_i\}$ are called *acoustic activity map*, below (Fig. 5).

3.3 Peak detection algorithm

We apply peak detection to the acoustic activity map at every time frame, yielding a set of Q peak observations $O = [\boldsymbol{o}_1 \dots \boldsymbol{o}_Q]$ with the corresponding acoustic activity peak values Γ_q with $q = 1 \dots Q$.

The peak detection algorithm is a greedy sequential algorithm similar to [15], however taking advantage of the continuous spherical harmonic DOA maps. We define the set of nodes $\boldsymbol{s}_i$ (cf. Sect. 2) on an equidistant three-dimensional grid with $i = 1 \dots G$ positions and a grid spacing of d_s, here chosen to be 0.25 m. The nodes are used to evaluate the acoustic activity following the procedure of equations (28)–(32). The maximum of these grid values

$$\Gamma_q = \max_i \gamma_i \quad \text{for } i = 1 \dots G \tag{33}$$

is selected. It being the global maximum of the acoustic activity, it likely is the loudest sound object in the scene.

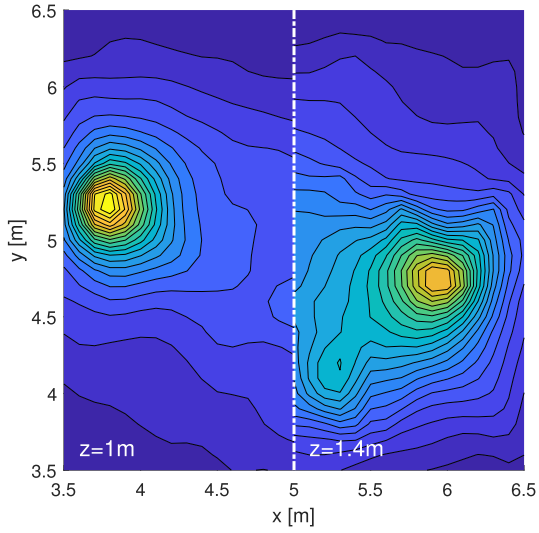


Figure 5. The acoustic activity map evaluated at equidistant grids on horizontal planes for visualization with two visible peaks.

The position of the peak, the observation o_q, is defined as the grid position corresponding to that maximum

$$\boldsymbol{o}_q = \boldsymbol{s}_j \quad \text{where } j = \arg\max_i \gamma_i \quad \text{for } i = 1 \dots G. \tag{34}$$

To detect local maxima representing other sound objects and at the same time avoid ghost peaks, we apply peak deletion on the DOA maps after each iteration of peak detection. This removal of peaks is similar to approaches [42, 43] for directional loudness editing in Ambisonics or generalized spherical beamforming [44–46]. A particularly suitable suppression function is denoted as

$$\mathcal{D}_p^{(q+1)} = \left(\mathbf{I} - \mathrm{diag}\{\boldsymbol{a}^{(q)}\}\boldsymbol{y}_N(\boldsymbol{\theta}_{0,p}^{(q)})\boldsymbol{y}_N(\boldsymbol{\theta}_{0,p}^{(q)})^\mathrm{T}\right)\mathcal{D}_p^{(q)} \tag{35}$$

and applied to each single-perspective DOA map $\mathcal{D}_p$. The function zeros the spherical harmonic representation at the direction $\boldsymbol{\theta}_0$, which is the perspective-to-peak direction

$$\boldsymbol{\theta}_{0,p}^{(q)} = \frac{\boldsymbol{o}_q - \boldsymbol{p}_p}{||\boldsymbol{o}_q - \boldsymbol{p}_p||}. \tag{36}$$

This removes an on-axis normalized, order-weighted beam pattern using

$$\boldsymbol{a}^{(q)} = \frac{\boldsymbol{a}}{\boldsymbol{y}_N(\boldsymbol{\theta}_{0,p}^{(q)})^T \mathrm{diag}\{\boldsymbol{a}\}\boldsymbol{y}_N(\boldsymbol{\theta}_{0,p}^{(q)})}, \tag{37}$$

that is directed towards the peak to be erased. The order weights $\boldsymbol{a}$ are essentially arbitrary, but maxRE-weights [33] of an order slightly smaller than the one of the DOA map proved to work well. This deletion step is done before continuing to detect the next loudest peak, as it minimizes the likelihood of an erroneous detection of ghost peaks associated with sub-optimal ray intersections belonging to the previous peak, cf. Figure 6a. A more elaborate explanation of the deletion function and order weights can be found in [47]. The sequential peak picking

is repeated until either a defined number Q of observations is reached or a peak value threshold is reached. The step sequence is presented in Algorithm 2.

Algorithm 2: Step sequence of the peak picking and deletion algorithm.

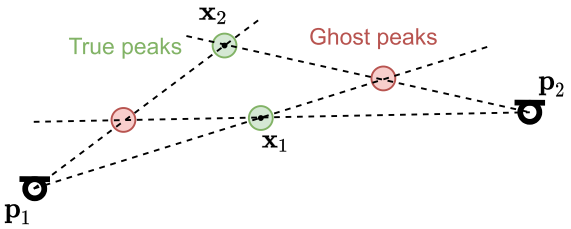
while *Peak magnitude is above threshold and maximum number of observations not reached* **do**
  (1) Compute acoustic activity map (32)
  (2) Pick global maximum (33 and grid position 34)
  (3) **for each** *Perspectives* **do**
    (a) Compute direction vector (36)
    (b) Apply peak deletion (35)
  end
end

3.4 Peak tracking with particle filters

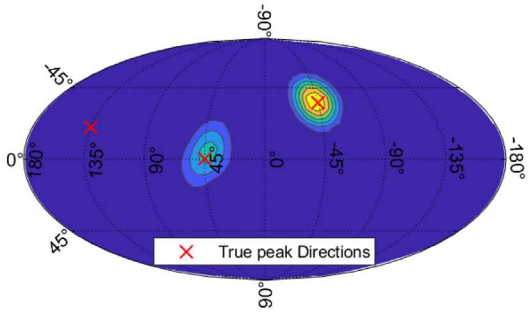
The peak observations $O = [\boldsymbol{o}_1 \ldots \boldsymbol{o}_Q]$ introduced in Section 3.3 are instantaneous for a time instant m. They are most likely noisy and therefore not directly useable for position estimation of the salient sound objects. To compute time-coherent trajectories for this position estimation, particle filters are introduced. Extending the concepts introduced in [21–23], particle filters are used in conjunction with a probabilistic detection algorithm, which evaluates the aforementioned peak observations using transitional probabilities. This evaluation involves a procedure to **(a)** start tracking a detected peak, **(b)** continue tracking a peak and **(c)** stop tracking a peak. Each peak considered valuable is associated with its own particle filter instance. Details on such an instance are found in Section 3.5. For all observations and known objects the procedure below is considered, starting at the most prominent peak:

(a) The lifecycle of a tracking instance starts when the value $P_q^{(m)}(\mathcal{H}_{\text{new}}) \in [0,1]$ for the peak observation o_q exceeds a threshold of 0.7. Is this the case, then a particle filter for this peak is initialized at the time instant m.
(b) The continuation of a currently tracked peak is determined by the value $P_s^{(m)} \in [0,1]$. If it exceeds a threshold of 0.6, then the sound object is deemed existent, active, and observable, and a location estimation is computed for the current time instant m. After first creation of an instance according to **(a)**, the threshold must be exceeded for longer than 0.1 s for the estimation to be considered viable, to avoid spurious detection.
(c) Complementing **(b)**: if $P_s^{(m)}$ falls below the threshold of 0.6 for ≥0.6 s, the peak is deemed vanished and the instance is discontinued.

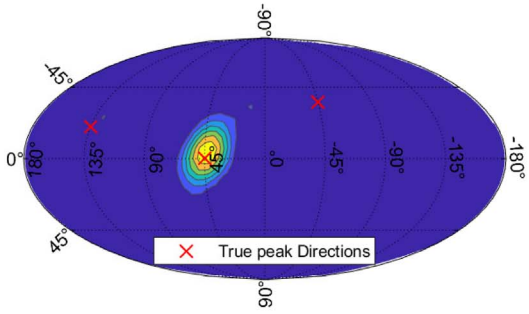
The procedure to compute $P_q^{(m)}(\mathcal{H}_{\text{new}})$ and $P_s^{(m)}$ is explained in Section 3.7, but it is valuable to describe the particle filter as applied in this work and its caveats in computation before.



(a) Intersection of directional information can lead to ghost peaks in three-dimensional data.



(b) DOA map before peak deletion. Maximum at (-45,-40)



(c) Renormalized DOA map after peak deletion. Maximum at (45,0)

Figure 6. (a) visualizes the problem of ghost peaks. (b) and (c) show the application of the peak deletion function, which removes directional components from the DOA maps $\mathcal{D}_p$. Intersection of directional information can lead to ghost peaks in three-dimensional data. DOA map before peak deletion. Maximum at (−45, −40). Renormalized DOA map after peak deletion. Maximum at (45, 0).

3.5 Particle filter dynamics

A particle filter is an estimation method employing a swarm of particles to estimate statistical measures. Here, it is used to estimate the continuous, true position of a peak in the activity map from the statistical centroid of the particle-swarm positions. The particle swarm random-samples the map around the true peak that it is directed to follow, and each particle has its own inertia and momentum. The concept of particle filters has been applied in estimation problems in audio applications such as in [19, 22, 23]. A non-exhaustive list of literature on particle filter theory is [48–55].

Since the peak observations O introduced in Section 3.3 are not time-coherent and their positional accuracy is

determined by the grid spacing, the particle filtering approach is used to obtain a consistent and continuous peak position estimation over time.

When a new peak is observed, as described in Section 3.4, an instance of a particle filter is initialized. This is done by sampling $N = 100$ positions following a multivariate Gaussian distribution

$$\boldsymbol{x}_i \sim \mathcal{N}\left(\boldsymbol{o}_q, \boldsymbol{\Sigma}_{\text{init}}\right), \tag{38}$$

centered around the peak observation $\boldsymbol{o}_q$ (34). The covariance matrix is chosen, so the sampling range covers the inaccuracies introduced by the grid spacing d_s and is defined as

$$\boldsymbol{\Sigma}_{\text{inst}} = \text{diag}\left\{\begin{bmatrix} d_{\text{grid}}^2 & d_{\text{grid}}^2 & d_{\text{grid}}^2 \end{bmatrix}\right\}. \tag{39}$$

A particle is now defined by a state vector

$$\boldsymbol{s}_i = \begin{bmatrix} \boldsymbol{x}_i \\ \dot{\boldsymbol{x}}_i \end{bmatrix}, \tag{40}$$

holding position and velocity, set to zero initially, in three dimensions, and a weight q_i determining the *importance*, as named in [55]. The particle weights q_i are the normalized acoustic activity map values

$$q_i = \frac{\gamma_i}{\sum_{j=1}^{N} \gamma_j} \quad \text{for } i = 1 \ldots N, \tag{41}$$

where γ_i are the acoustic map values for the particle positions $\boldsymbol{x}_i$ following the procedure from equations (28) to (32) without peak deletion.

The estimation of peak position is done by taking a weighted mean, or centroid,

$$\hat{\boldsymbol{x}} = \sum_{i=1}^{N} \boldsymbol{x}_i q_i \tag{42}$$

of the particle positions, see (40).

The velocity values are not required above, but ensure that the trajectories obtained for the peak positions evolve smoothly over time. To ensure trajectories can be adjusted to reasonable physical qualities, the excitation-damping dynamical model adapted from [19] and [22, 23] is applied to each particle $\boldsymbol{s}_i$. The prediction step

$$\boldsymbol{s}_i^{(m)} = \mathbf{M}\boldsymbol{s}_i^{(m-1)} + \mathbf{F}^{(m)}. \tag{43}$$

computes the particle state for time instant m from the state of the previous time instant $m + 1$. For this, the state-space system from [19] with the time step Δt determined by hopsize is denotes as the system matrix

$$\mathbf{M} = \begin{bmatrix} 1 & 0 & 0 & \Delta t & 0 & 0 \\ 0 & 1 & 0 & 0 & \Delta t & 0 \\ 0 & 0 & 1 & 0 & 0 & \Delta t \\ 0 & 0 & 0 & a_{\text{dyn}} & 0 & 0 \\ 0 & 0 & 0 & 0 & a_{\text{dyn}} & 0 \\ 0 & 0 & 0 & 0 & 0 & a_{\text{dyn}} \end{bmatrix}, \tag{44}$$

as well as the process noise

$$\mathbf{F}^{(m)} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\text{pr}}) \tag{45}$$

with

$$\boldsymbol{\Sigma}_{\text{pr}} = \text{diag}\{[\,0 \quad 0 \quad 0 \quad b_{\text{dyn}} \quad b_{\text{dyn}} \quad b_{\text{dyn}}\,]\}. \tag{46}$$

Including the process noise $\mathbf{F}$ is necessary to account for the unpredictability in the trajectory of unknown moving objects. The two factors

$$a_{\text{dyn}} = \mathrm{e}^{-\alpha_{\text{dyn}}\Delta t}, \tag{47}$$

$$b_{\text{dyn}} = \beta_{\text{dyn}}\sqrt{1 - a_{\text{dyn}}^2}, \tag{48}$$

determine the behavior of the particles. The factors α_{dyn} and β_{dyn} are chosen dependent on the expected object movement. They determine the damping of velocity and the process noise influence on the prediction step. In this application, these values were chosen to be $\alpha_{\text{dyn}} = 2$ and $\beta_{\text{dyn}} = 0.04$ as the active sound objects are mostly static.

The particle filter dynamics above describes how the particles move, however as the process noise lead to random directions, the particle swarm still needs to be led towards the locations of the greatest acoustic activity. This achieved by *importance re-sampling*, which is a step that relocates particles of low importance to locations of particles with high importance; a relocated particle continues its motion from its new location. In this work this re-sampling is applied every time instant m, subsequent to the application of the object detection algorithm, which is subsequent to particle position prediction, cf. Algorithm 1. The re-sampling scheme used in this algorithm is the *systematic* approach introduced in [56, 53]; other methods are *multinomial* [48, 53], *residual* [56, 52], and *stratified* ([50], is done by sampling).

3.6 Anti-causal computation

Since the probabilistic calculations and certain time thresholds of the algorithm impose a lag on the detection of actively trackable sound objects, anti-causal look-ahead processing is introduced. After the first analysis of the microphone array recordings, a part of the process is repeated in the time-reversed direction.

Each object detection event from the causal process is used as starting point for the anti-causal processing. The probabilistic detection algorithm is applied in the same manner as in the causal process, however without the ability to initialize new track instances. Only the continuation or removal of existing instances as introduced in Section 3.3 is applied. The particle filter prediction step itself is applied using a negative time step size $\Delta t \to -\Delta t$ in (44).

The computation of the prediction step, particle weights, and position estimation is only done for time instants m at which the causal processing has not declared the object as active yet.

3.7 Object detection algorithm

Section 3.3 already introduced the concepts of threshold values determining the initialization, continuation, and removal of tracking instances. These values are computed by a probabilistic algorithm introduced in [21, 22] that is adapted to our three-dimensional application.

The sequential peak detection algorithm introduced in Section 3.3 yields Q instantaneous observations $\boldsymbol{o}_q$ (34) and peak activity values Γ_q (33). These peaks can be caused by objects directly, but are still strongly affected by noise, reflections, discretization artifacts, and other interference. Therefore, each detected peak (observation) is evaluated in terms of its viability for the peak tracking algorithm. To accomplish this, the probability of the observation (detected peak) to existing sound objects needs to be quantified, or the likelihood of it belonging to a new object, or it being a false detection. The local relative peak prominence of each peak observation is the relation between first and therefore maximum peak $q = 1$ and the remaining peaks $q > 1$

$$P_q = \frac{\Gamma_q}{\Gamma_1}, \tag{49}$$

in contrast to empirically defined calculation in [22] or histogram peak values in [23] that are found in literature.

For the subsequent explanation, we assume an active sound scene of $s = 1 \ldots S$ initialized peak tracking instances yielding the position estimates $\hat{\boldsymbol{x}}_s$ (42). All variables introduced in Section 3.5 now carry the additional index for the tracking instance s. For each sound-object track s, we define the observability as in [22, 23], but denoted differently as

$$O_s^{(m)} = A_s^{(m)} E_s^{(m)}. \tag{50}$$

for the time instant m. The object *activity* $A^{(m)}$ and the object *existence* $E^{(m)}$ are probability values for the object peaks that are being actively tracked. The activity is computed using the first-order Markov model

$$A_s^{(m)} = p_{\mathrm{A}} \tilde{A}_s^{(m)} + p_{\bar{\mathrm{A}}} \left[1 - \tilde{A}_s^{(m)}\right] \tag{51}$$

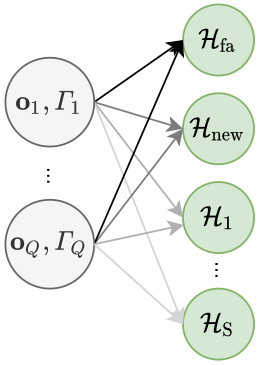
with the transition probabilities $p_{\mathrm{A}} = 0.95$ and $p_{\bar{\mathrm{A}}} = 0.05$ representing the state transition from active to active state and inactive to active state, respectively, just as in [22]. The intermediate activity value $\tilde{A}_s^{(m)}$ in (51) is the probabilistic combination

$$\tilde{A}_s^{(m)} = \left[1 + \frac{\left(1 - A_s^{(m)}\right)\left(1 - P_s^{(m-1)}\right) + \epsilon}{A_s^{(m)} P_s^{(m-1)} + \epsilon}\right]^{-1} \tag{52}$$

including the object tracking probability P_{s}. The small value $\epsilon > 0$ added to the denominator ensures numerical robustness.

Then, the recursive existence is defined in [22, 23] as the function

$$E^{(m)} = P_s^{(m-1)} + \left(1 - P_s^{(m-1)}\right) \frac{\mu E^{(m-1)}}{1 - \mu E^{(m-1)}}, \tag{53}$$



(a) Observation-hypothesis mapping.

	q	
r	1	2
1	$\mathcal{H}_{\mathrm{fa}}$	$\mathcal{H}_{\mathrm{fa}}$
2	$\mathcal{H}_{\mathrm{fa}}$	$\mathcal{H}_{\mathrm{new}}$
3	$\mathcal{H}_{\mathrm{fa}}$	$\mathcal{H}_1$
4	$\mathcal{H}_{\mathrm{new}}$	$\mathcal{H}_{\mathrm{fa}}$
5	$\mathcal{H}_{\mathrm{new}}$	$\mathcal{H}_{\mathrm{new}}$
6	$\mathcal{H}_{\mathrm{new}}$	$\mathcal{H}_1$
7	$\mathcal{H}_1$	$\mathcal{H}_{\mathrm{fa}}$
8	$\mathcal{H}_1$	$\mathcal{H}_{\mathrm{new}}$
9	$\mathcal{H}_1$	$\mathcal{H}_1$

(b) Example with Q=2 and S=1.

Figure 7. (a) Graphical visualization of possible mapping combinations; (b) with $Q = 2$ and $S = 1$, a total number of nine possible combinations exist. The selection introduced with δ_r, selects all r where the hypothesis is included, i.e. $\delta_{r,\mathrm{fa}}$ would lead to the set of $r \in \{1, 2, 3, 4, 7\}$.

that automatically assumes high values of high tracking probability $P_s^{(m-1)}$, or maps small values of $P_s^{(m-1)}$ with smoothed falling slopes in $E^{(m)}$, with the smoothing constant μ, which is set to 0.5.

The paper [22] introduces the following hypotheses: the observation with index q is caused (1) by a tracked sound object with index $s = 1 \ldots S$, denoted as $\mathcal{H}_s$, (2) by a false detection caused by interference, denoted as $\mathcal{H}_{\mathrm{fa}}$, or (3) by a new sound object, denoted as $\mathcal{H}_{\mathrm{new}}$.

The set of observations $\boldsymbol{o}_q$ with $q = 1 \ldots Q$ has to be mapped to these hypotheses. This results in $r = 1 \ldots (S + 2)^Q$ mapping combinations (cf. Fig. 7), which have to be evaluated. Assuming conditional independence of the single observations, this is done by computing the association probabilities

$$p(r|O^{(m)}) = \prod_{q=1}^{Q} p(\boldsymbol{o}_q^{(m)}|\mathcal{H}_q) p(\mathcal{H}_q) \tag{54}$$

for each combination r. The term $\mathcal{H}_q$ is the hypothesis mapped to the observation q for the combination r. The term $p(\boldsymbol{o}_q^{(m)}|\mathcal{H})$ describes the likelihood of an observation at the position $\boldsymbol{o}_q$ while mapped to a certain hypothesis. It is defined as

$$p(\boldsymbol{o}_q^{(m)}|\mathcal{H}_q) = \begin{cases} p_{\mathrm{fa}}\left(\boldsymbol{o}_q^{(m)}\right) & \text{if } \mathcal{H}_q = \mathcal{H}_{\mathrm{fa}} \\ p_{\mathrm{new}}\left(\boldsymbol{o}_q^{(m)}\right) & \text{if } \mathcal{H}_q = \mathcal{H}_{\mathrm{new}} \\ p(\boldsymbol{o}_q^{(m)}|\hat{\boldsymbol{x}}_s^{(m-1)}) & \text{if } \mathcal{H}_q = \mathcal{H}_s \end{cases}. \tag{55}$$

This involves *a priori* knowledge. The three-dimensional spatial distribution function $p_{\mathrm{fa}}(\boldsymbol{o}_q)$ describes the probability of an observation $\boldsymbol{o}$ being a false detection, e.g. volumes blocked by obstacles will have a high probability for false detection. Similarly, $p_{\mathrm{new}}(\boldsymbol{o}_{\mathrm{q}})$ describes the known probability of new sound objects appearing in the scene, e.g. on a

stage. Both can be set to a uniform distribution if there is no specific knowledge is available. The function $p(\boldsymbol{o}_q^{(m)}|\hat{\boldsymbol{x}}_s^{(m-1)})$ is not two dimensional as in [22, 23], but defined as a three-dimensional multivariate Gaussian distribution

$$p(\boldsymbol{o}_q^{(m)}|\hat{\boldsymbol{x}}_s^{(m-1)}) = \mathcal{N}\big(\boldsymbol{o}_q, \hat{\boldsymbol{x}}_s, \Sigma_s\big) \tag{56}$$

centered at the latest estimate $\hat{\boldsymbol{x}}_s$ and evaluated at the observation position $\boldsymbol{o}_q$. The covariance is estimated by the particle positions of the corresponding particle filter as

$$\Sigma_\mathrm{s} = c\,\mathrm{Cov}\{\boldsymbol{x}_{s,i}\}, \tag{57}$$

with a scaling, chosen to be $c = 4$, here.

The second term

$$p(\mathcal{H}_q) = \begin{cases} \mathrm{p_{fa}}\,(1-P_q) & \text{if } \mathcal{H}_q = \mathcal{H}_\mathrm{fa} \\ \mathrm{p_{new}}\,P_q & \text{if } \mathcal{H}_q = \mathcal{H}_\mathrm{new} \\ P_q\,O^{(m)} & \text{if } \mathcal{H}_q = \mathcal{H}_s \end{cases} \tag{58}$$

describes probabilities incorporating the peak prominence (49) and time instant-wise computation of the observability (50). It additionally introduces the factors $p_\mathrm{fa} = 0.8$ and $p_\mathrm{new} = 0.2$ for tuning the impact of the prominence P_q.

The subsequent summation for all combinations containing the hypothesis gives the marginal probabilities of observation-hypothesis mappings. This is computed as

$$\begin{gathered} P_q^{(m)}(\mathcal{H}) = \sum_r \delta_{r,\mathcal{H}}\, p\big(r|O^{(m)}\big), \\ \text{with } \mathcal{H} \in \{\mathcal{H}_\mathrm{fa}, \mathcal{H}_\mathrm{new}, \mathcal{H}_{s=1}, \ldots, \mathcal{H}_{s=S}\} \end{gathered} \tag{59}$$

where the Kronecker delta denotes the selection from the set of $(S+2)^Q$ mappings where the hypothesis $\mathcal{H}$ is included (see Fig. 7). The value $P_q^{(m)}(\mathcal{H}_\mathrm{new})$ is required for tracking control in Section 3.4. The object tracking probability of each known sound object is evaluated as

$$P_s^{(m)} = \sum_{q=1}^{Q} P_q^{(m)}(\mathcal{H}_s), \tag{60}$$

and is also employed in tracking control of Section 3.4.

4 Technical evaluation of object position estimation

The accuracy of sound object detection and positional accuracy was evaluated by simulating sound scene recordings with varying noise interference [47]. The simulated scene was created with the simulation library for MATLAB library *MCRoomSim*[2] [57]. The room is a 6 m × 6 m × 3.5 m box model with default room acoustic settings. For this evaluation, four static sound objects, two male and two female speakers from the *EBU SQAM* collection [58], are situated in the simulated room at static positions (cf. Tab. 1). Microphone arrays are located at 1 m intervals ranging from 3.5 m to 6.5 m on the x and y coordinate. To evaluate the detection capabilities also covering grids with volumetric extent, vertical layers are positioned between 0.5 m and 2.5 m with 1 m spacing along the z axis. This results in 48 virtual microphone arrays. The simulation models microphone arrays as non-coincident using the array geometry of the Oktava MK4012 [37]. The scenes point of origin is located at the center of the room's floor.

Table 1. The sound object positions and *EBU SQAM* collection track numbers. The first 5 s of the signals are used in the simulation.

No.	x [m]	y [m]	z [m]	*EBU SQAM* Nr.
1	5.25	5.50	1.75	49 (0–5 s)
2	6.00	4.75	1.50	50 (0–5 s)
3	5.25	4.00	1.50	51 (0–5 s)
4	3.75	5.25	1.00	52 (0–5 s)

To measure the accuracy of the algorithm, the simulated scene is analyzed at different levels of noise interference. The SNR here is defined as the relation between the loudest microphone signal and uncorrelated white noise added to all microphone signals independently with the same signal energy. The SNR values for this evaluations were SNR $\in \{9, 12, 15, 18\}$ dB.

4.1 Error measures

Mean Distance Error: The measure is defined as the distance between the ground truth to the nearest sound object. Only 1-to-1 mappings are allowed, therefore if an object is already in use for calculation then the next-nearest will be used if existent. The measure is computed for every sound object $j = 1 \ldots 4$ and averaged over N_trial trials and M time frames whenever actively tracked sound objects were found:

$$\mathrm{MDE}_j = \frac{1}{N_\mathrm{trial}} \sum_{n=1}^{N_\mathrm{trial}} \frac{1}{M} \sum_{m=1}^{M} \min_i \, ||\boldsymbol{s}_{j,\mathrm{true}}^{(m,n)} - \hat{\boldsymbol{s}}_i^{(m,n)}||. \tag{61}$$

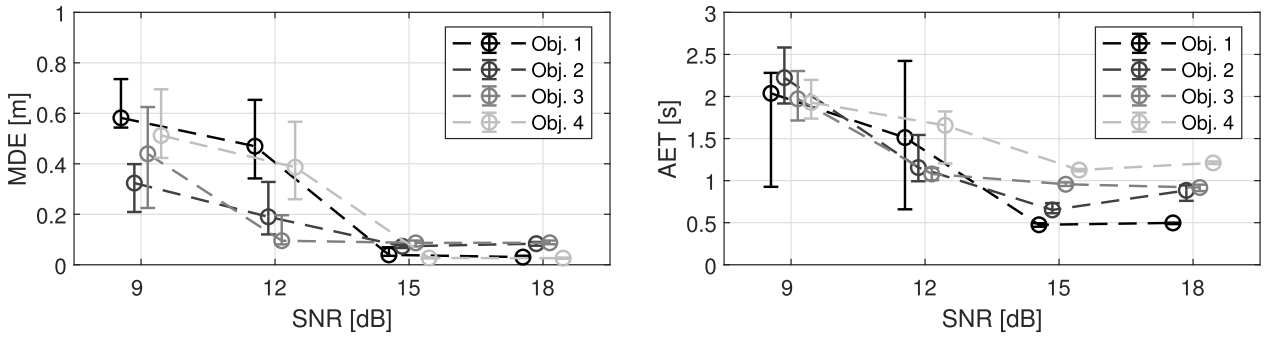
Activity Error Time: The sound object activity is the time where false positives or false negatives are observed in sound object activity. This is done by summation of time frame lengths where there is no nearest sound object existent or the sound object probability differs from 1. It is a strict measure, for which the resulting error time can be high despite getting accurate results. Here it is used to show the relative improvement with varying noise interference. It is defined as

$$\mathrm{AET}_j = \sum_{m=1}^{M} \phi_{j,\mathrm{active}}(m)\Delta t, \tag{62}$$

$$\phi_{j,\mathrm{active}}(m) = \begin{cases} A_{j,\mathrm{truth}}^{(m)} & \text{(a)} \\ |P_s^{(m)} - A_{j,\mathrm{truth}}^{(m)}| & \text{(b)} \end{cases}. \tag{63}$$

Case (a) is applied if no nearest sound object could be found that was not assigned yet, while (b) applies whenever the

[2] Library available at https://github.com/Andronicus1000/MCRoomSim.



(a) MDE (b) AET

Figure 8. (a) Higher SNR values in scene recordings yield good positional accuracy in object localization between 2 and 10 cm. (b) The AET of the sound objects decreases with higher SNR. At lower SNRs, the confidence intervals suggest again a strong variation in results indicative of unstable measurement results. Shown are mean, and 95% confidence intervals.

tracked sound object s is nearest to the jth truth value still available. The true value $A^{(m)}_{j,\mathrm{truth}}$ is defined by manual labeling based on the instantaneous magnitude of the sound object.

4.2 Results

The mean distance error (MDE) is low with SNR values at 15 dB and 18 dB, as visualized in Figure 8a. The confidence intervals suggest stable values between 2 and 10 centimeters. On the other hand, large confidence intervals at the smaller SNR values suggest unreliable detection and position estimation. The activity error time (AET), see Figure 8b, shows a similar dependency on SNR confirming the SNR requirement for this system at $\geq$15 dB.

In summary, the evaluation showed promising accuracy that led us to confide in rendering that might be able to synthesize a spatial auditory image of detected sound objects that closely resembles the one of the recorded scene.

5 Perceptual evaluation

To evaluate the effectiveness of the proposed procedure of rendering, listening experiments have been conducted. A simple scene consisting of two static active sound objects is analysed. The male speaker (Track Nr. 50) and the piano (Track Nr. 60) of the *EBU SQAM* collection [58]. Here the time interval [1;8] s (piano) and [3;8] s (speech) are used where start of the speech signal is 2 s delayed behind the piano signal. A second scene is simulated where only one of the objects is present, which in turn is used as the baseline for the rendering using the estimated position data from the two-object scene. The purpose of this approach is to include possible artifacts and inaccuracies stemming from inter-object interference in analysis whilst minimizing complexity for the listeners of the experiments.

The experiment was conducted in two parts. First, static listener perspectives were presented to the listener where the virtual listener is positioned at fixed locations facing the active sound object. The second part presented dynamic perspectives, representing a linear motion of the listener perspective along a pre-defined path. In this case the look direction was varied between four distinct orientations. These pre-defined modes of motion are used for auralization using a binaural decoder, so that the experiment can be conducted using headphones without the need for head/position tracking equipment.

The methodology of the experiment was a MUSHRA-like [59] comparative evaluation asking for the perceptual similarity to the reference, also described as the *authenticity*. The conditions of the comparison include the simulated reference, the proposed method, two broadband rendering methods, and an anchor.[3]

The *reference* is the binaural rendering of the listener perspective as simulated by MCRoomSim [57].

The *proposed* condition is a binaural rendering following the procedure introduced in Section 2 and Section 3.

The *VLO* approach was introduced in [30, 31] and is a broadband spatial rendering method to enable acoustic scene playback with spatially distributed surround recordings. Refer to Figure 3a for an overview. The virtual loudspeaker objects are encoded in third-order Ambisonics and decoded to binaural signals with the IEM BinauralDecoder [60].

The Vector-Based Intensity Panning (VBIP) approach is a simple superposition of three surround recordings transformed to first-order Ambisonics signals weighted with the areal coordinate approach. Again, the IEM BinauralDecoder[4] was used for decoding to headphone signals. A previous study on the performance of *VLO* and *VBIP* as done in [61].

To give a low-rating reference, a mono version of the VBIP condition was added as *anchor*. This was visually not identifiable and part of the randomly sorted multistimulus presentation.

5.1 Experiment

Both the static and dynamic listening tasks used a room measuring 6 m × 6 m × 3.5 m simulated in MCRoomSim with the same grid layout as in the technical evaluation

[3] Conditions available at https://phaidra.kug.ac.at/o:107325.
[4] Plugin available at https://plugins.iem.at/.

above: equally spaced according to Figure 9 and vertically repeated in three layers at 0.5 m, 1.5 m and 2.5 m height. The simulated arrays model the geometry of the Oktava MK4012 [37].

The static-perspective tasks evaluated four listener positions that are visualized in Figure 9. Such a task consisted of a the comparative rating of authenticity where the reference was always visible to the listener and the order of conditions was randomized and not visually identifiable. Each static position was rated twice by each participant, also randomly sequenced within the set of multi-stimulus tasks.

Further the dynamic-perspective tasks consisted of the comparative rating of authenticity with visible reference. The four look directions A, B, C, D shown in Figure 9 were evaluated twice by each participant, in randomized condition and task orders.

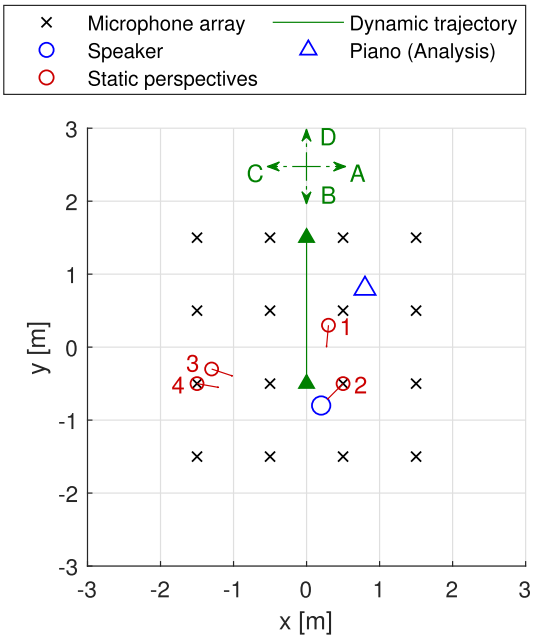
The signals were optimized for the most common AKG and Beyerdynamic high end models to minimize coloration. In total, 16 expert listeners aged between 24 and 39 (average age: 29) took part in the listening experiment taking 30 min on average to complete it. The pairwise statistical significance was assessed with a Wilcoxon signed rank test [62] with Bonferroni-Holm correction [63]. There were 32 responses for each condition yielding $4 \times 32 = 128$ responses when merging over positions 1 to 4 in part 1 and look directions A–D in part 2. The data proved to be consistent enough to be merged.

All plots in Figure 10 show the sample median of collected sample populations and ≥95% confidence intervals. The choice of the median over the mean is based on its higher robustness towards outliers, especially relatively small sample populations. The confidence intervals are computed by applying the binomial test [64, 65] to the samples of each evaluated condition.

5.2 Results

Static-perspective tasks: The single position ratings are visualized in Figure 10a (Median and ≥95% confidence intervals). Participants rated the *proposed* approach higher than all the other conditions, with statistical significance ($p < 0.001$) at positions 1–3. At position 4, however, *VLO* rendering shows significantly higher rating ($p < 0.001$) than *VLO* at 1–3, and no significant difference to the *proposed* one ($p = 0.0541$). The comparison of merged *VLO* data from Positions 1 and 3 on the one hand and 2 and 4 on the other hand (not displayed) exhibits an advantage ($p < 0.001$) of the direct perspective positions over the interpolated ones. The ratings of the *VBIP* approach show a decrease with distance, when ratings are compared with locations in Figure 9. The difference is significant when comparing the farthest and closest position ($p < 0.001$).

Dynamic-perspective tasks: Figure 10c shows, the ratings for all conditions are very similar over look directions A–D suggesting that look direction is of little influence. Moreover, the *proposed* and *VLO* methods are consistently rated higher ($p < 0.001$) than the *VBIP* condition. Between *proposed* and *VLO*, the advantage is not as strong but



(a) Perspective positions 1-3, dynamic path with look directions A,B,C,D.

	x [m]	y [m]	z [m]
Speaker	0.2	-0.8	1.5
Piano	0.8	0.8	1.75
Pos 1	0.3	0.3	1.5
Pos 2	0.5	-0.5	1.5
Pos 3	-1.3	-0.3	1.5
Pos 4	-1.5	-0.5	1.5
Path start	0	-0.5	1.5
Path end	0	1.5	1.5

(b) Position coordinates.

Figure 9. The perspectives used in the listening experiment. (a) The static perspective positions and the trajectory of the dynamic perspective used in the listening experiment. (b) Lists coordinates of static perspective positions and path start and end point.

existent at all look direction A ($p = 0.0696$), B ($p = 0.0218$), C ($p < 0.001$) or D ($p = 0.0650$). This experiment supports the findings of [61] where the VLO approach performed better than the VBIP approach.

The merged responses across all directions of the dynamic-perspective experiment (Fig. 10d) imply a significant mean difference ($p < 0.001$) between ratings of *proposed* and *VLO*/*VBIP*, supporting the results of the static-perspective experiment (Fig. 10b).

5.3 Discussion

This listening evaluation confirmed the intended improvements in scene resynthesis by comparison with

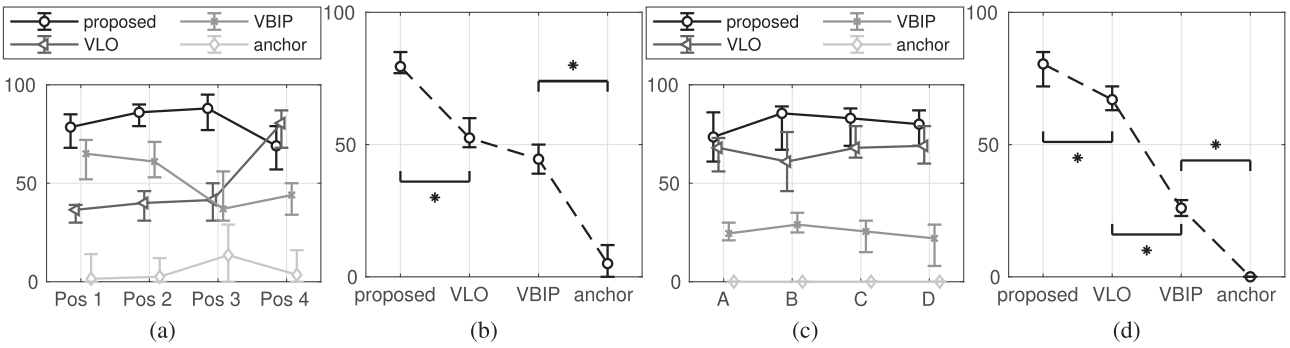


Figure 10. The results of the second experiment comparing dynamic perspectives for (a) single look direction data and conditions as well as the (b) merged-look-direction data. Pictured are the sample median, the ≥95% confidence intervals and notable statistical significance is marked.

established broadband methods. Despite the limited sample size, most of the statistical results were significant. The static and dynamic-perspective parts of the experiment indicated a significant increase in authenticity of the proposed methods over the compared ones.

For the static-perspective part, the authenticity, i.e. the similarity to the reference, of the *proposed* is consistently rated high, and in the majority of the cases higher than the alternatives. As the one exception, the observed drop off at position 4 is most likely due to the high rating of *VLO* and combined with the limited scale of the ratings. Further, *VLO* shows ratings seemingly depending on the listener position and the significant increase in rating at position 4 is due to the fact that this position is a direct microphone array position and far away from a source. There, the surround perspective of the microphone position provides accurate reproduction just as with *VBIP*, however providing a better room impression owed to the rich diversity of VLOs and their directions. By contrast, spatial reproduction with *VLO* and *VBIP* suffers at interpolated, less diffuse listener perspectives.

The dynamic-perspective part of the experiment shows an increase in ratings for *VLO* as the interpolation is perceived smoother and is rated almost as high as *proposed*. As residual signals of *proposed* are based on the *VLO* approach, smoothness and general impression are understandably similar whenever the virtual listener has moved away from the sources. The advantages lie in auditory localization of direct sounds through the proposed object direct signal extraction and encoding. The data shows this improvement in the single-direction as well as the merged-data comparisons.

6 Conclusion

In this contribution we proposed an analysis and resynthesis method for acoustic scenes recorded with distributed surround microphone arrays, in the investigated case tetrahedral A-format Ambisonics microphones. We could show that multi-perspective recordings provide sufficiently much additional information for rendering with significantly improved spatial accuracy and authenticity, already when performed broadband, in the time domain, only. This effectively avoids any risk of introducing musical-noise artifacts that any potentially more effective time-frequency processing intrinsically bears.

A numerical experiment considered sound objects of a simulated scene and could prove good accuracy in object position and signal activity estimation, and it revealed a 15 dB SNR or direct to diffuse ratio limit that local microphones around the active sound object should be able to satisfy. We could further verify the suspected improvement compared to two known broadband perspective interpolation approaches in a two-part listening experiment; its results show improvements in authenticity and spatial definition.

References

1. T. Pihlajamäki, V. Pulkki: Projecting simulated or recorded spatial sound onto 3d-surfaces, in AES Conference: 45th International Conference: Applications of Time-Frequency Processing in Audio, 03 2012. Available: http://www.aes.org/e-lib/browse.cfm?elib=16198.
2. T. Pihlajamäki, V. Pulkki: Synthesis of complex sound scenes with transformation of recorded spatial sound in virtual reality. Journal of the Audio Engineering Society 63 (2015) 542–551. Available: http://www.aes.org/e-lib/browse.cfm?elib=17840.
3. V. Pulkki: Directional audio coding in spatial sound reproduction and stereo upmixing, in AES Conference: 28th International Conference: The Future of Audio Technology – Surround and Beyond, 06, 2006. Available: http://www.aes.org/e-lib/browse.cfm?elib=13847.
4. V. Pulkki, A. Politis, M.-V. Laitinen, J. Vilkamo, J. Ahonen: First-order directional audio coding (dirac). Parametric Time-Frequency Domain Spatial Audio 10 (2017) 89–140. https://doi.org/10.1002/9781119252634.ch5.
5. A. Plinge, S.J. Schlecht, O. Thiergart, T. Robotham, O. Rummukainen, E.A.P. Habets: Six-degrees-of-freedom binaural audio reproduction of first-order ambisonics with distance information, in: AES International Conference on Audio for Virtual and Augmented Reality 082018). Available: http://www.aes.org/e-lib/browse.cfm?elib=19684.
6. N. Barrett, S. Berge: A new method for b-format to binaural transcoding, in Audio Engineering Society Conference: 40th International Conference: Spatial Audio: Sense the Sound of

Space, 10, 2010. Available: http://www.aes.org/e-lib/browse.cfm?elib=15527.
7. E. Stein, M.M. Goodwin: Ambisonics depth extensions for six degrees of freedom, in AES Conference: 2019 AES International Conference on Headphone Technology, 08 2019, Available: http://www.aes.org/e-lib/browse.cfm?elib=20514.
8. A. Allen, B. Kleijn: Ambisonic soundfield navigation using directional decomposition and path distance estimation, in ICSA, Graz, Austria, 09 2017.
9. M. Kentgens, A. Behler, P. Jax: Translation of a higher order ambisonics sound scene based on parametric decomposition, in IEEE ICASSP (2020) 151–155. https://doi.org/10.1109/ICASSP40776.2020.9054414.
10. L. Birnie, T. Abhayapala, P. Samarasinghe, V. Tourbabin: Sound field translation methods for binaural reproduction, in IEEE WASPAA (2019) 140–144. Available: https://doi.org/10.1109/WASPAA.2019.8937274.
11. E. Bates, H. O'Dwyer, K.-P. Flachsbarth, F.M. Boland: A recording technique for 6 degrees of freedom VR, in AES Convention, Vol. 144. Audio Engineering Society, 05 2018. Available: http://www.aes.org/e-lib/browse.cfm?elib=19418
12. H. Lee: A new multichannel microphone technique for effective perspective control, in AES Convention, Vol. 140. Audio Engineering Society, 05 2011. Available: https://www.aes.org/e-lib/browse.cfm?elib=15804.
13. A. Brutti, M. Omologo, P. Svaizer: Localization of multiple speakers based on a two step acoustic map analysis. IEEE ICASSP (2008) 4349–4352. Available: https://doi.org/10.1109/ICASSP.2008.4518618.
14. A. Brutti, M. Omologo, P. Svaizer: Multiple source localization based on acoustic map de-emphasis. EURASIP Journal on Audio, Speech, and Music Processing 2010 (2010). 147495. https://doi.org/10.1155/2010/147495.
15. P. Hack, Multiple source localization with distributed tetrahedral microphone arrays. Master's Thesis, Institute of Electronic Music and Acoustics, University of Music and Performing Arts Graz, Graz, Austria, 2015. Available: http://phaidra.kug.ac.at/o:12797
16. G. Del Galdo, O. Thiergart, T. Weller, E.A. Habets: Generating virtual microphone signals using geometrical information gathered by distributed arrays, in 2011 Joint Workshop on Hands-free Speech Communication and Microphone Arrays, IEEE, 05 2011. Available: https://doi.org/10.1109.
17. O. Thiergart, G. Del Galdo, M. Taseska, E.A.P. Habets: Geometry-based spatial sound acquisition using distributed microphone arrays. IEEE Transactions on Audio, Speech, and Language Processing 21 (2013) 2583–2594. https://doi.org/10.1109/TASL.2013.2280210.
18. X. Zheng: Soundfield navigation: Separation, compression- and transmission. Ph.D. Dissertation, University of Wollongong, 2013. Available: https://ro.uow.edu.au/theses/3943/.
19. D.B. Ward, E.A. Lehmann, R.C. Williamson: Particle filtering algorithms for tracking an acoustic source in a reverberant environment. IEEE Transactions on Speech and Audio Processing 11 (2003) 11 https://doi.org/10.1109/TSA.2003.818112.
20. M.F. Fallon, S.J. Godsill: Acoustic source localization and tracking of a time-varying number of speakers. IEEE Transactions on Audio, Speech, and Language Processing 20 (2012) 1409–1415. https://doi.org/10.1109/TASL.2011.2178402.
21. J.-M. Valin, F. Michaud, J. Rouat: Robust 3D localization and tracking of sound sources using beamforming and particle filtering. IEEE ICASSP (2006). https://doi.org/10.1109/ICASSP.2006.1661100.
22. J.-M. Valin, F. Michaud, J. Rouat: Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering. Elsevier Science 55 (2007) 216–228. Available: https://arxiv.org/pdf/1602.08139.pdf.
23. S. Kitić, A. Guérin: Tramp: Tracking by a real-time ambisonic-based particle filter, in LOCATA Challenge Workshop, 09 2018. Available: https://arxiv.org/abs/1810.04080.
24. J.G. Tylka, E. Choueiri. Soundfield navigation using an array of higher-order ambisonics microphones, in AES International Conference on Audio for Virtual and Augmented Reality, 09 (2016). Available: http://www.aes.org/e-lib/browse.cfm?elib=18502.
25. J.G. Tylka, E.Y. Choueiri: Domains of practical applicability for parametric interpolation methods for virtual sound field navigation. Journal of the Audio Engineering Society 67 (2019) 882–893. Available: http://www.aes.org/e-lib/browse.cfm?elib=20702.
26. J.G. Tylka: Virtual navigation of ambisonics-encoded sound fields containing near-field sources. PhD dissertation, Princeton University, 2019. Available: http://arks.princeton.edu/ark:/88435/dsp011544br958.
27. N. Mariette, B.F.G. Katz, K. Boussetta, O. Guillerminet: Sounddelta: A study of audio augmented reality using wifi-distributed ambisonic cell rendering. AES Convention, Vol. 128. Audio Engineering Society, 2010. Available: http://www.aes.org/e-lib/browse.cfm?elib=15420.
28. C. Schörkhuber, R. Höldrich, F. Zotter: Triplet-based variable-perspective (6DoF) audio rendering from simultaneous surround recordings taken at multiple perspectives, in Fortschritte der Akustik (DAGA), Hannover, Germany, 04 2020. Available: https://pub.dega-akustik.de/DAGA_2020/data/articles/000295.pdf.
29. E. Patricio, A. Rumiński, A. Kuklasiński, L. Januszkiewicz, T. Żernicki: Toward six degrees of freedom audio recording and playback using multiple ambisonics sound fields. AES Convention, Vol. 146, Audio Engineering Society, 2019. Available: http://www.aes.org/e-lib/browse.cfm?elib=20274.
30. P. Grosche, F. Zotter, C. Schörkhuber, M. Frank, R. Höldrich: Method and apparatus for acoustic scene playback. Patent WO2018077379A1 (2018). Available: https://patents.google.com/patent/WO2018077379A1.
31. F. Zotter, M. Frank, C. Schörkhuber, R. Höldrich: Signal-independent approach to variable-perspective (6DoF) audio rendering from simultaneous surround recordings taken at multiple perspectives, in Fortschritte der Akustik (DAGA), Hannover, Germany. 04 2020. Available: https://pub.dega-akustik.de/DAGA_2020/data/articles/000458.pdf.
32. D. Rivas Méndez, C. Armstrong, J. Stubbs, M. Stiles, G. Kearney: Practical recording techniques for music production with six-degrees of freedom virtual reality. AES Convention, Vol. 145, Audio Engineering Society, 2015. Available: http://www.aes.org/e-lib/browse.cfm?elib=19729.
33. F. Zotter, M. Frank: Ambisonics, 1st edn., Vol. 19 ofSpringer Topics in Signal Processing, Springer International Publishing, 2019. https://doi.org/10.1007/978-3-030-17207-7.
34. A. Politis: Microphone array processing for parametric spatial audio techniques. PhD dissertation, Aalto University, 2016. Available: http://urn.fi/URN:ISBN:978-952-60-7037-7.
35. J. Ivanic, K. Ruedenberg: Rotation matrices for real spherical harmonics. direct determination by recursion. The Journal of Physical Chemistry 100 (1996) 6342–6347. https://doi.org/10.1021/jp953350u.
36. C. Schörkhuber, M. Zaunschirm, R. Höldrich: Binaural rendering of ambisonic signals via magnitude least squares, in Fortschritte der der Akustik (DAGA), Munich, Germany,

03 2018. Available: https://pub.dega-akustik.de/DAGA_2018/data/articles/000301.pdf.
37. Oktava GmbH: Oktava mk-4012 (2019). Available: http://www.oktava-shop.com/images/product_images/popup_images/4012.jpg.
38. E. Hille: Analytic Function Theory, 2nd edn., Vol. 1. Chelsea Publishing Company, New York, 1982.
39. A. Politis, S. Delikaris-Manias, V. Pulkki: Direction-of-arrival and diffuseness estimation above spatial aliasing for symmetrical directional microphone arrays. IEEE ICASSP (2015) 6–10. https://doi.org/10.1109/ICASSP.2015.7177921.
40. T. Wilding: System parameter estimation of acoustic scenes using first order microphones, Master's thesis. Institute of Electronic Music and Acoustics, University of Music and Performing Arts Graz, Graz, Austria, 2016. Available: http://phaidra.kug.ac.at/o:40685.
41. Z. He, A. Cichocki, S. Xie, K. Choi: Detecting the number of clusters in n-way probabilistic clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (2010) 2006–2021. https://doi.org/10.1109/TPAMI.2010.15.
42. M. Kronlachner: Spatial transformations for the alteration of ambisonic recordings. Master's thesis (2014). Available: http://phaidra.kug.ac.at/o:8569.
43. M. Hafsati, N. Epain, J. Daniel: Editing ambisonics sound scenes. ICSA, Graz, Austria, 09 2017.
44. M. Jeffet, B. Rafaely: Study of a generalized spherical array beamformer with adjustable binaural reproduction (2014) 77–81. https://doi.org/10.1109/HSCMA.2014.6843255.
45. N. Shabtai, B. Rafaely: Generalized spherical array beamforming for binaural speech reproduction. IEEE/ACM Transactions on Audio, Speech, and Language Processing 22 (2014) 238–247. https://doi.org/10.1109/TASLP.2013.2290499.
46. M. Jeffet, N. Shabtai, B. Rafaely: Theory and perceptual evaluation of the binaural reproduction and beamforming tradeoff in the generalized spherical array beamformer. IEEE/ACM Transactions on Audio, Speech, and Language Processing 24 (2016) 708–718. https://doi.org/10.1109/TASLP.2016.2522649.
47. M. Blochberger: Multi-perspective scene analysis from tetrahedral microphone recordings. Master's thesis (2020). Available: https://phaidra.kug.ac.at/o:104549.
48. B. Efron, R. Tibshirani: An Introduction to the Bootstrap, Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1994. Available: https://books.google.at/books?id=gLlpIUxRntoC.
49. J.S. Liu, R. Chen: Blind deconvolution via sequential imputations. Journal of the American Statistical Association 90 (1995) 567–576. https://doi.org/10.1080/01621459.1995.10476549.
50. P. Fearnhead: Sequential monte carlo methods in filter theory. PhD dissertation, University of Oxford, 1998.
51. G. Kitagawa: Monte carlo filter and smoother for non-gaussian nonlinear state space models. Journal of Computational and Graphical Statistics 5 (1996) 1–25. https://doi.org/10.1080/10618600.1996.10474692.
52. J. Liu, R. Chen: Sequential Monte Carlo methods for dynamic systems. Journal of the American Statistical Association 93 (1998) 1032–1044. https://doi.org/10.1080/01621459.1998.10473765.
53. J. Carpenter, P. Clifford, P. Fearnhead: An improved particle filter for non-linear problems. IEE Proceedings Radar Sonar and Navigation 146 (1999) 2–7. https://doi.org/10.1049/ip-rsn:19990255.
54. A. Doucet, N. de Freitas, N. Gordon: Sequential Monte Carlo Methods in Practice, 1st edn., Information Science and Statistics. Springer-Verlag, New York, 2001. https://doi.org/10.1007/978-1-4757-3437-9.
55. S. Särkkä: Bayesian Filtering and Smoothing, Institute of Mathematical Statistics Textbooks. Cambridge University Press, 2013. https://doi.org/10.1017/CBO9781139344203.
56. D. Whitley: A genetic algorithm tutorial. Statistics and Computing 4 (1994) 65–85. https://doi.org/10.1007/BF00175354.
57. A. Wabnitz, N. Epain, C. Jin, A. van Schaik: Room acoustics simulation for multichannel microphone arrays. ISRA, Melbourne, Australia, 08 2010. Available: https://www.acoustics.asn.au/conference_proceedings/ICA2010/cdrom-ISRA2010/Papers/P5d.pdf.
58. EBU: Sound Quality Assessment Material recordings for subjective tests, 2008. Available: https://tech.ebu.ch/publications/sqamcd.
59. ITU, ITU-R BS.1534-3: Method for the subjective assessment of intermediate quality level of audio systems, 2015. Available: https://www.itu.int/rec/R-REC-BS.1534-3-201510-I.
60. D. Rudrich: IEM Plugin Suite. IEM, 2019. Available: https://plugins.iem.at/.
61. D. Rudrich, F. Zotter, M. Frank: Evaluation of interactive localization in virtual acoustic scenes. Fortschritte der Akustik (DAGA), Kiel, Germany, 09 2017. Available: https://pub.dega-akustik.de/DAGA_2017/data/articles/000182.pdf.
62. F. Wilcoxon: Individual comparisons by ranking methods. Biometrics Bulletin 1 (1945) 80–83. Available: http://www.jstor.org/stable/3001968.
63. S. Holm: A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6 (1979) 65–70. Available: http://www.jstor.org/stable/4615733.
64. D. Altman, D. Machin, T. Bryant, M. Gardner: Statistics with Confidence, Confidence Intervals and Statistical Guidelines, 2nd edn., BMJ Books (2000).
65. M. Eid, M. Gollwitzer, M. Schmitt: Statistik und Forschungsmethoden, 5th edn. Julius Beltz, 2017.

Cite this article as: Blochberger M & Zotter F. 2021. Particle-filter tracking of sounds for frequency-independent 3D audio rendering from distributed B-format recordings. Acta Acustica, **5**, 20.