Audio Engineering Society

Conference Paper 13

Presented at the AES 5th International Conference on Audio for Virtual and Augmented Reality
2024 August 19–21, Redmond, WA, USA

This paper was peer-reviewed as a complete manuscript for presentation at this conference. This paper is available in the AES E-Library (http://www.aes.org/e-lib), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

6DoF Rendering of Multi-Point Ambisonics Recordings at Listening Positions Exterior to the Area Defined by Microphone Positions

Jussi Leppänen[1], Archontis Politis[2], Mikko-Ville Laitinen[1], Lauros Pajunen[1], and Antti Eronen[1]

[1]*Nokia Technologies, Finland*
[2]*Audio Research Group, Tampere University, Finland*

Correspondence should be addressed to Jussi Leppänen (jussi.leppanen@nokia.com)

ABSTRACT

This work presents a method for 6DoF binaural rendering from multi-point first- or higher-order Ambisonic (FOA/HOA) recordings for the case when the listener is outside of the area covered by the microphones. This work builds on top of a previously presented method for 6DoF binaural rendering when the listener is inside the area covered by the microphones [1]. The method measures spatial parameters in the time-frequency domain that summarize the scene at the microphone positions. It then estimates the same parameters at the listener position through spatial interpolation of the measured ones. In this work, the spatial parameter calculation is modified to accommodate situations where spatial interpolation is not possible as described in the previous work. A real-time 6DoF listening preference test demonstrates that the proposed method provides plausible rendering when the listener is outside of the microphone area. This work is related to the upcoming MPEG-I immersive audio standard (ISO/IEC 23090-4) which is defining 6DoF immersive audio technology for AR and VR scenarios.

1 Introduction

Media technologies are rapidly evolving from standard video recording and transmission with stereophonic audio to fully immersive experiences, often offering some degree of interactivity and supported by a new generation of wearable devices such as head-mounted displays (HMDs). Such immersive media technologies are exemplified, e.g. in the upcoming MPEG-I media standard. In terms of immersive audio, the standard defines technology for 6DoF rendering of objects, channels and Ambisonics for Virtual Reality (VR) and Augmented Reality (AR) scenarios. The Committee Draft of the MPEG-I immersive audio standard has been recently published [2]. Furthermore, telecommunications in general are also moving quickly to immersive experiences, as exemplified for example by the recent 3GPP IVAS audio standard [3]. Spatial audio capturing, coding, transmission, and reproduction, with a special focus on binaural headphone delivery, is a core component of all those emerging standards. An important part of it is 6DoF rendering of immersive sound scenes that adapt the processing dynamically to the position and head-orientation of the user to deliver the appropriate spatial audio perspective. Such 6DoF rendering is more straightforward when audio objects

are rendered with virtual acoustics, but less so when it involves rendering of real recorded sound scenes. Such a task of rendering a 6DoF binaural perspective of a sound scene captured with multiple Ambisonic microphones is specified in MPEG-I immersive audio, and is the focus of this study.

There have been various approaches to this problem proposed in literature. For a comprehensive review please refer to [4]. A rough categorization can be attempted based on some key directions. A class of methods focuses on simple broadband mixing of the recordings [5, 6, 7] aiming for simplicity and high signal fidelity, but without necessarily accurate spatial cue delivery. Another class of methods attacks the problem through the intermediate task of sound field interpolation [8, 9, 10, 11, 12], based on physical sound field models and their expansions to plane waves or spherical waves. Such approaches have a solid theoretical foundation but their performance is typically limited to low frequencies for a practical number of microphones. Finally, a third class of methods uses source localization and spatial filtering principles to turn the recordings into an object-based representation of distinct time-varying source components with estimated positions and some residual ambience, which can subsequently be rendered spatially at arbitrary positions [4, 13, 14, 15]. Such approaches have been demonstrated with high perceptual performance in real-time [4], but they require well defined scenes of a few distinct sources in reverberant conditions to capture their parameters accurately with a small number of microphones.

Additional capabilities that are of interest are whether the methods can support rendering of sources in the interior of the region covered from the microphones (interior sources), or only sources at its exterior (exterior sources). The third class of methods can handle both cases, while sound field interpolation methods can only handle exterior sources. Another capability, and the focus of this work, is whether the method can provide plausible rendering at positions outside of the interior region (exterior rendering). The third class of methods can achieve that due to their object-based representation of the scene, while sound field interpolation methods can naturally extrapolate to a limited range from the interior region. However, none of the above methods can handle arbitrary complex scenes of interior and exterior sources at reverberant environments with high perceived spatial quality, while being practical enough to operate with a low computational load and with a small number of microphones covering a wide 6DoF area.

An attempt to such an approach was proposed by Politis et al. [1] describing a method for 6DoF binaural rendering from multi-point first- or higher-order Ambisonic (FOA/HOA) recordings. The proposed method was the winner of the multiple-Ambisonics rendering task of the Call-for-Proposals of the MPEG-I 6DoF immersive audio [16] and is part of the upcoming standard (ISO/IEC 23090-4). The proposal uses as a basis Directional Audio Coding with optimal mixing (OM-DirAC), which is a parametric, signal dependent method for 3DoF rendering of Ambisonics signals [17, 18]. In the method the input (prototype) FOA/HOA signal is adapted to match its binaural cues to that of a target spatial covariance matrix (SCM). The target SCM is obtained from spatial parameters calculated from the input HOA/FOA signal.

In [1], 6DoF rendering is achieved by first calculating a set of spatial parameters $\mathscr{P}^{(m)}$ at each of the microphone positions $\mathbf{P}_M = [\mathbf{p}_1, ..., \mathbf{p}_m]$, and then interpolating them to estimate the spatial parameters at the listener position ($\mathscr{P}^{(\mathrm{L})}$), see Figure 1. Interpolation is done for microphones surrounding the listener, derived via triangulation of the microphone positions. The target SCM is then calculated from the interpolated spatial parameters and the closest FOA/HOA signal to the listener is used as the prototype signal. Finally, the output binaural signal is obtained by adapting the protype signal to have the same binaural cues as the target SCM.

This work concentrates on 6DoF rendering when spatial metadata interpolation as described in Politis et al. [1] is not possible, i.e. when the listener is outside of the triangulated microphone area. We propose an approach where the spatial parameters are first interpolated at the edge of the microphone area and then modified based on a projection radius and the listener distance from the microphone area edge.

In the following, we first briefly go through the processing presented in Politis et al. [1], in Section 2 and then explain the proposed method in Section 3. Subjective evaluation of the proposed method is described in Section 4.

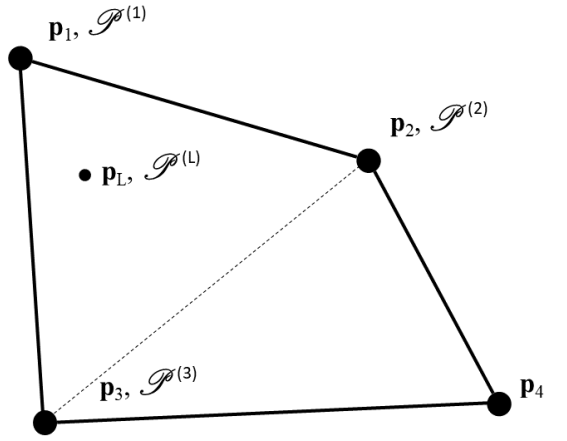


Fig. 1: The listener is inside a triangle defined by microphones at positions $[\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3]$. For each of these, a set of spatial parameters are calculated ($[\mathscr{P}^{(1)}, \mathscr{P}^{(2)}, \mathscr{P}^{(3)}]$). Spatial parameters at the listener position, $\mathscr{P}^{(L)}$, are then obtained by weighted interpolation of the spatial parameters at the microphone positions.

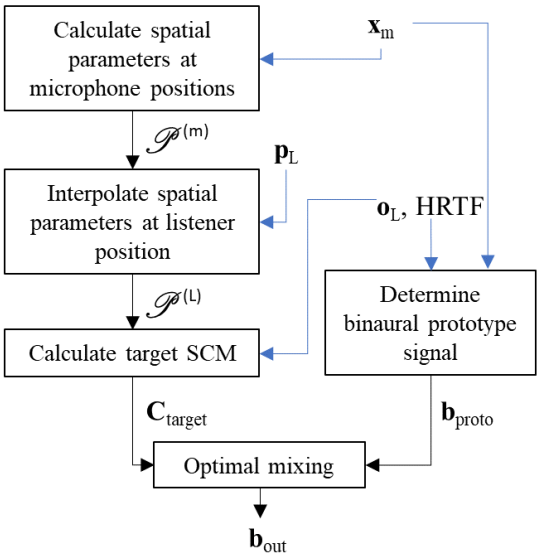


Fig. 2: Block diagram of the main steps in [1].

2 6DoF rendering of multi-point Ambisonic recordings

In this section, we describe the main steps of the method presented in [1]. Spatial parameter calculation and interpolation is covered as well as obtaining the target SCM and prototype signal from which the final binaural output is obtained. Figure 2 shows a block diagram of the main steps of the method.

2.1 Spatial parameter calculation

The spatial parameters for each microphone are calculated from their associated time-frequency domain FOA/HOA signals for each frequency bin and frame index. The calculated spatial parameters $\mathscr{P}^{(m)}$ comprise energy E_m, intensity $\mathbf{i}_m$ and diffuseness ψ_m and are calculated as follows for each frame index n and frequency band f:

$$E_m(f,n) = \frac{1}{2} \sum_{c \in \{1..4\}} [\mathbf{x}_m(f,n)]_c \cdot \overline{[\mathbf{x}_m(f,n)]_c} \quad (1)$$

$$\mathbf{i}_m(f,n) = \Re\big([\,[\mathbf{x}_m(f,n)]_1 \cdot \overline{[\mathbf{x}_m(f,n)]_4}, \; [\mathbf{x}_m(f,n)]_1 \cdot \overline{[\mathbf{x}_m(f,n)]_2}, \; [\mathbf{x}_m(f,n)]_1 \cdot \overline{[\mathbf{x}_m(f,n)]_3}\,]^{\mathrm{T}}\big) \quad (2)$$

$$\psi_m(f,n) = 1 - ||\mathbf{i}_m(f,n)||/E_m(f,n) \quad (3)$$

where $[\mathbf{x}_m(f,n)]_c$ is the cth channel of the time-frequency domain FOA/HOA input signal for microphone m and $\overline{[\mathbf{x}_m(f,n)]_c}$ is the complex conjugate of $[\mathbf{x}_m(f,n)]_c$. $\Re(\cdot)$ is the real part of complex scalars or vectors.

2.2 Spatial parameter interpolation

After deriving spatial parameters at the microphone locations, the spatial parameters at the listener location may be derived as explained in [1]. In brief, the scheme involves Delauney triangulation of the microphone positions $\mathbf{p}_m$ as a pre-processing step and then during the rendering, based on the triangulation, calculating Barycentric interpolation weights for the microphones. This scheme has the benefit of only requiring spatial parameters to be calculated for microphones comprising the triangle that the listener is located in as the interpolation weights for the other microphones are zero. Also, for further practical benefit, arbitrary microphone arrangements are possible to be used.

To determine the interpolation weights, first, the set of microphones comprising the triangle the user is in ($m \in \{i,j,k\}$) is determined. Then, for these microphones, interpolation weights are obtained as follows:

$$\begin{bmatrix} g_i(n) \\ g_j(n) \end{bmatrix} = [\mathbf{p}_i - \mathbf{p}_k, \mathbf{p}_j - \mathbf{p}_k]^{-1} \cdot (\mathbf{p}_L(n) - \mathbf{p}_k), \\ g_k(n) = 1 - g_i(n) - g_j(n). \tag{4}$$

The weights fulfil $\sum_{m\in\{i,j,k\}} g_m = 1$ and are zero for all other microphones. Based on those, the estimated spatial parameters at the listener position are

$$\tilde{E}_{\mathrm{L}}(f,n) = \sum_{m\in\{i,j,k\}} g_m(n) E_m(f,n) \tag{5}$$

$$\tilde{\mathbf{i}}_{\mathrm{L}}(f,n) = \sum_{m\in\{i,j,k\}} g_m(n) \big(\mathbf{i}_m(f,n)/E_m(f,n)\big) \tag{6}$$

$$\tilde{\psi}_{\mathrm{L}}(f,n) = 1 - ||\tilde{\mathbf{i}}_L(f,n)||. \tag{7}$$

2.3 Binaural output

The target SCM $\mathbf{C}_{\mathrm{target}} \in \mathbb{C}^{2\times 2}$ is calculated from the interpolated spatial parameters as shown below

$$\mathbf{C}_{\mathrm{target}}(f,n) = \tilde{E}_L(f,n)\Big(\tilde{\psi}_{\mathrm{L}}(f,n)\mathbf{F}(f) \\ + (1-\tilde{\psi}_{\mathrm{L}}(f,n))\tilde{\mathbf{h}}(f,n)\tilde{\mathbf{h}}^{\mathrm{H}}(f,n)\Big), \tag{8}$$

where $\mathbf{F} \in \mathbb{C}^{2\times 2}$ is the binaural diffuse field coherence matrix which can be constructed from the HRTF set; see e.g. [19]. $\tilde{\mathbf{h}}(f,n)$ is the HRTF for the direction of arrival $\mathbf{u}_{\mathrm{DOA}}(f,n)$ pointed by the intensity vector $\mathbf{i}(f,n)$.

A binaural prototype signal is then obtained from the closest microphone to the listener:

$$\mathbf{b}_{\mathrm{proto}}(f,n) = \\ \mathbf{D}_{\mathrm{ambi2bin}}(f) \cdot \mathbf{R}_{\mathrm{rot,sh}}(\mathbf{o}_{\mathrm{L}}) \cdot \mathbf{x}_{m_{\mathrm{close}}}(f,n), \tag{9}$$

where $\mathbf{D}_{\mathrm{ambi2bin}}(f)$ is ambisonics to binaural conversion matrix for frequency band f calculated using the *magnitude least-squares* approach [20] and $\mathbf{R}_{\mathrm{rot,sh}}(\mathbf{o}_{\mathrm{L}}) \in \mathbb{R}^{(L+1)^2\times(L+1)^2}$ is a spherical harmonic domain rotation matrix for real spherical harmonics derived based on the listener head orientation $\mathbf{o}_{\mathrm{L}}$ as e.g. in [21].

At this stage of the processing, we have a binaural prototype signal and a target SCM. The SCM is calculated based on the interpolated spatial parameters and incorporates all binaural cues of interest, namely the binaural spectrum, inter-aural level and phase differences, and inter-aural coherence at the listener position. The final binaural output is obtained by adapting the prototype signal to have the same binaural cues as the target SCM.

The final output is obtained by:

$$\mathbf{b}_{\mathrm{out}}(f,n) = \mathbf{M}(f,n)\mathbf{b}_{\mathrm{proto}}(f,n), \tag{10}$$

where $\mathbf{M}(f,n)$ is a mixing matrix derived according to [18]. The mixing matrix is derived such that the following property is satisfied:

$$\mathbf{M}(f,n)\mathbf{C}_{\mathrm{proto}}(f,n)\mathbf{M}^{\mathrm{H}}(f,n) = \mathbf{C}_{\mathrm{target}}(f,n), \tag{11}$$

where $\mathbf{C}_{\mathrm{proto}}(f,n)$ is a SCM obtained from $\mathbf{b}_L(f,n)$:

$$\mathbf{C}_{\mathrm{proto}}(f,n) = \mathbb{E}[\mathbf{b}_L(f,n)\mathbf{b}_L^{\mathrm{H}}(f,n)], \tag{12}$$

where $\mathbb{E}[\cdot]$ denotes statistical expectation.

3 Exterior Rendering

The above description assumes that the listener is inside a triangle defined by the triangulation (microphone area). When the listener moves outside of the area defined by the triangles, deriving the interpolation weights for the interpolation of the energy-normalized intensity vectors via barycentric weights is not possible. However, we would not like to restrict the listener from moving outside of the microphone area. Furthermore, the listener should experience plausible 6DoF rendering that reacts to their movements. Figure 3 illustrates the problem in more detail with the help of an example.

Since interpolation as described above is not possible, a new scheme for estimating spatial parameters for listener positions outside the microphone area is required. We shall refer to this as exterior rendering in the following text, see Figure 4. The spatial parameters should react to user movement in a meaningful manner and there should be no discontinuities as the user moves from inside of the microphone area to the outside. In addition, the scheme should work for cases where both interior and exterior sound sources are present.

Based on the requirements discussed above, a scheme for exterior rendering was chosen where spatial parameters are first calculated for a projected (to the edge of the microphone area) listener position $\mathbf{p}_{\mathrm{L,proj}}$ and then these parameters are modified based on the listener distance from the edge of the microphone area. The modifications are done such that when the listener

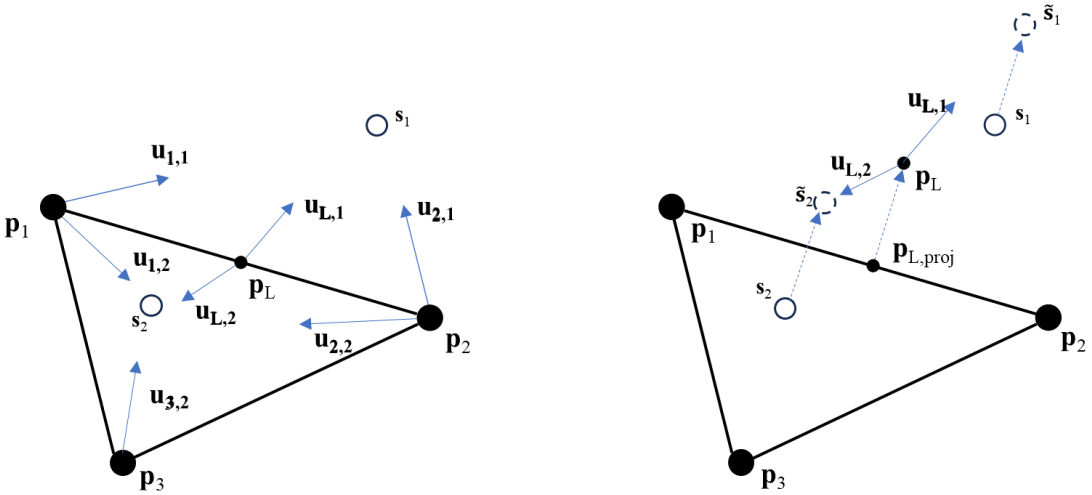


Fig. 3: On the left, a listener is at the edge of the microphone area defined by microphones at positions $\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3$. An external sound source is at position $\mathbf{s}_1$ and an internal sound source is at position $\mathbf{s}_2$. Spatial parameters have been calculated at the microphone positions and then interpolated at the listener position $\mathbf{p}_\mathrm{L}$ (here we show only the DOAs of the intensity vectors ($\mathbf{u}$) for clarity). The interpolated DOA vectors point towards $\mathbf{s}_1$ and $\mathbf{s}_2$ and the listener hears the sound sources from the correct directions. On the right, the listener has moved away from the microphone area. Since the listener is not inside a triangle anymore, interpolation as described in Section 2 is not possible. If we were to interpolate spatial parameters based on the closest position of the microphone area to the listener ($\mathbf{p}_\mathrm{L,proj}$), for example, a mismatch in perceived positions of the sound sources ($\tilde{\mathbf{s}}_1$ and $\tilde{\mathbf{s}}_2$) and the correct positions ($\mathbf{s}_1$ and $\mathbf{s}_2$) would occur. The sound sources would appear as following the listener.

position is located on the microphone area edge, the modified spatial parameters obtained from exterior rendering are equal to those obtained by interior rendering. The modifications affect the direction-of-arrival (DOA) of the intensity vector as well as the diffuseness. Figure 5 shows a block diagram of the main steps of the proposed method. Note that the method is formulated for HOA microphones positioned at the same height, with planar interpolation of parameters. The method can be extended to 3D recording arrangements, using volumetric tetrahedral interpolation for parameter interpolation [22] instead of triangular one, and projection of the listener position onto surfaces of the 3D convex hull of the microphone positions instead of onto edges of the 2D convex hull in the present 2D case. However, planar recording arrangements are considered more common in practice and only the 2D case is presented herein. This limitation does not limit 6DoF rendering to planar DOAs, which could be analyzed at the full sphere, but it limits their interpolation to use only measurements on the same plane.

3.1 DOA modification

For a proper 6DoF experience, the sound source directions perceived by the user should change based on user movement. In the approach used here, we achieve this by mapping sound sources outside of the microphone area at the intersection of the DOA parameters $\mathbf{u}_\mathrm{DOA}(f,n)$ and a sphere with a radius of R_proj. The center of the sphere is placed at the position $\mathbf{p}_\mathrm{L,proj}$ on the edge of the microphone area that is closest to the listener position $\mathbf{p}_\mathrm{L}(n)$. The motivation for mapping sound sources onto the sphere is based on the assumption that the sources of interest are likely to be at some finite distance from the microphone area. The radius of the sphere can be set based on knowledge of the recorded scene. In some cases, we know, for example, the size of a room where the scene is recorded or perhaps even the locations of the sound sources outside of the microphone area and we can thus set the radius accordingly.

The original DOA parameter extracted from the in-

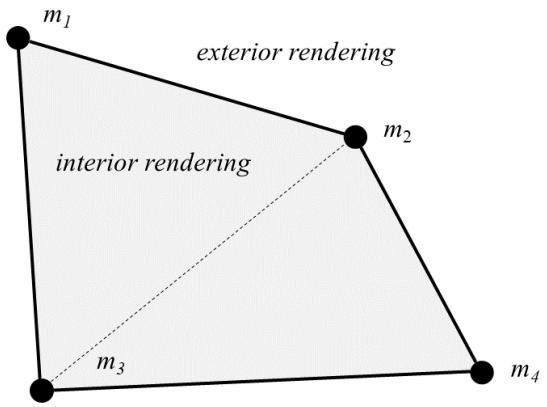


Fig. 4: When the listener is inside the area defined by the microphones, $m_{1...4}$ in this example, rendering is performed as described in Section 2. This shall be referred to as *interior rendering* in the following text. When the listener is outside of the area defined by the microphones, the proposed scheme is used (Section 3). This shall be referred to as *exterior rendering* in the following text.

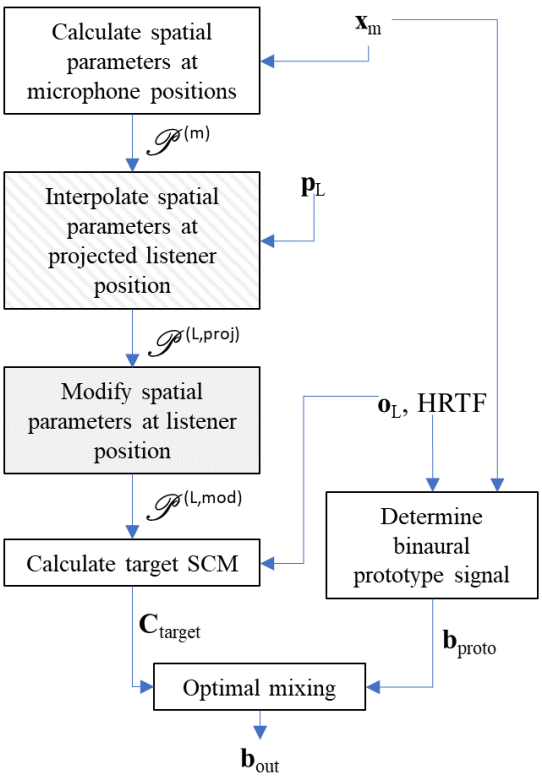


Fig. 5: Block diagram of the proposed method for exterior rendering. The filled in blocks are new or modified blocks compared to the block diagram in Figure 2.

tensity vector of each frequency band, $\mathbf{u}_{\mathrm{DOA}}(f,n)$, is modified to obtain a modified DOA, $\mathbf{u}_{\mathrm{DOA,mod}}(f,n)$, as shown in Figure 6. As the listener moves away from the microphone area, the interpolated DOA parameter, $\mathbf{u}_{\mathrm{DOA}}(f,n)$, is unchanged, but the modified one, $\mathbf{u}_{\mathrm{DOA,mod}}(f,n)$ is tracking the projected source position, hence, giving an impression of sources that react more realistically to motion.

3.2 Diffuseness modification

The DOA modification causes sound sources to be mapped to positions on the sphere and thus react to user movement in a 6DoF manner. However, the sound sources outside of the microphone area are rendered from plausible, but not necessarily correct, locations. Since the mapped sound source locations do not exactly match the sound source locations during capture, a modification (increase) of the diffuseness parameter is done when the listener approaches the mapped sound source positions. This strategy is used to avoid listener confusion because of perceived localization errors due to the mismatch between correct and mapped sound source locations. The mismatch would be most noticeable when the listener is close to the mapped source position, especially if the listener sees a visual representation of the sound source in a VR environment. The modified diffuseness parameter, $\tilde{\psi}_{\mathrm{L,mod}}(f,n)$, is obtained as shown in Equation 13. The diffuseness is increased as the distance between the listener and mapped sound source gets smaller.

$$\tilde{\psi}_{\mathrm{L,mod}}(f,n) = \max\left(\tilde{\psi}_{\mathrm{L}}(f,n), 1 - \frac{|\mathbf{u}_{\mathrm{DOA,mod}}(f,n)|}{R_{\mathrm{proj}}}\right) \tag{13}$$

3.3 Directional weighting

Directional weighting is applied to more heavily modify parameters from sound sources outside the capturing region, while applying less modifications to sources

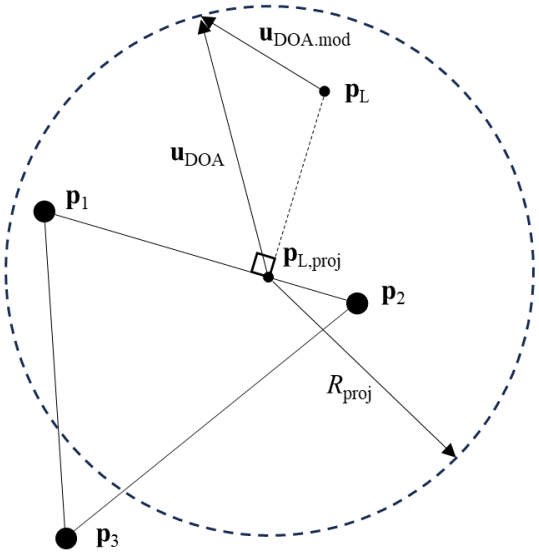


Fig. 6: During exterior rendering, a modified DOA vector, $\mathbf{u}_{\mathrm{DOA,mod}}(f,n)$, is obtained by first projecting the original DOA vector, $\mathbf{u}_{\mathrm{DOA}}(f,n)$, on to a point on a sphere with radius R_{proj} and then obtaining a vector from the current listener position $\mathbf{p}_{\mathrm{L}}(n)$ to the projected point.

inside the microphone area. The logic behind this strategy is that time-frequency bins that are dominated by sources from inside the interior rendering region, as indicated by their DOAs pointing towards the interior instead of away from it, should be mostly unaffected by the parameter modifications of Sections 3.1 - 3.2. The final modified DOA and diffuseness parameters are obtained using the directional weighting as follows:

$$\begin{aligned}\mathbf{u}_{\mathrm{DOA,final}}(f,n) = (1-w_{\mathrm{dir}}(f,n))\mathbf{u}_{\mathrm{DOA}} \\ + w_{\mathrm{dir}}(f,n)\mathbf{u}_{\mathrm{DOA,mod}}(f,n)\end{aligned} \tag{14}$$

$$\begin{aligned}\tilde{\psi}_{\mathrm{L,final}}(f,n) = (1-w_{\mathrm{dir}}(f,n))\tilde{\psi}_{L} \\ + w_{\mathrm{dir}}(f,n)\tilde{\psi}_{\mathrm{L,mod}}(f,n)\end{aligned} \tag{15}$$

where

$$w_{\mathrm{dir}}(f,n) = \frac{1}{2^{w_s}}(1+\mathbf{n}_{\mathrm{L}}(f,n)\cdot\hat{\mathbf{u}}_{\mathrm{DOA}}(f,n))^{w_s} \tag{16}$$

where $\mathbf{n}_{\mathrm{L}}$ is a unit vector pointing from $\mathbf{p}_{\mathrm{L,proj}}(n)$ to $\mathbf{p}_{\mathrm{L}}(n)$ (i.e. away from the microphone area) and $\hat{\mathbf{u}}_{\mathrm{DOA}}(f,n)$ is a unit vector pointing in the same direction as $\mathbf{u}_{\mathrm{DOA}}(f,n)$. $w_{\mathrm{dir}}(f,n)$ will be 1, when $\mathbf{n}_{\mathrm{L}}$ and $\hat{\mathbf{u}}_{\mathrm{DOA}}(f,n)$ point in the same direction. This means that when a DOA points away from the microphone area, the modified spatial parameters will be used as the final ones in Equations 14 and 15. As $\hat{\mathbf{u}}_{\mathrm{DOA}}(f,n)$ turns towards the microphone area, $w_{\mathrm{dir}}(f,n)$ decreases towards 0 giving more weight to the original spatial parameter values. w_s is a weighting sharpness, controlling how sharply w_{dir} decreases towards 0 as the angle between $\mathbf{n}_{\mathrm{L}}$ and $\hat{\mathbf{u}}_{\mathrm{DOA}}(f,n)$ increases.

3.4 Binaural output

The target SCM, $\mathbf{C}_{\mathrm{target}}$, is then calculated by Equation 8 using the obtained $\mathbf{u}_{\mathrm{DOA,final}}(f,n)$ and $\tilde{\psi}_{\mathrm{L,final}}(f,n)$. Otherwise the exterior rendering processing follows the description in Section 2.

4 Evaluation

The perceptual performance of the method was tested with a Multiple-stimulus Category Rating (MuSCR) test proposed specifically for evaluation in MPEG-I [23]. For the test, two versions of a synthesized scene of multiple simulated ideal third-order HOA microphones a few meters apart was used. Both scenes included 4 omnidirectional sources representing individual instruments in a music segment. In the first of the two scenes (labeled '*5 HOA*'), all sources were positioned inside a microphone area defined by 5 HOA microphones. Figure 7 shows the arrangement of the HOA microphones and sources. The scene was reverberated with a shoebox image source simulator, embedded in a $9 \times 10 \times 3$ m virtual room with an RT60 of 280 msec at 1 kHz. For both scenes, the exterior projection radius R_{proj} was set to 1 meter. The second scene (labeled '*3 HOA*') is identical to the first scene, except that it has two of the HOA microphones removed. Removing two of the HOA microphones causes three of the four sources to be positioned outside of the microphone area.

Evaluation was conducted using a real-time VR testing platform, the Audio Evaluation Platform (AEP) of MPEG-I [23], where listeners could switch freely between test items and rate them on a floating GUI appearing inside the VR scene, on a scale of 0 to 100. The visuals, rendered using Unity and delivered to the listeners on an HTC ViVe head-mounted display (HMD), comprised a simple visual representation of the room and loudspeaker graphic objects indicating the source positions.

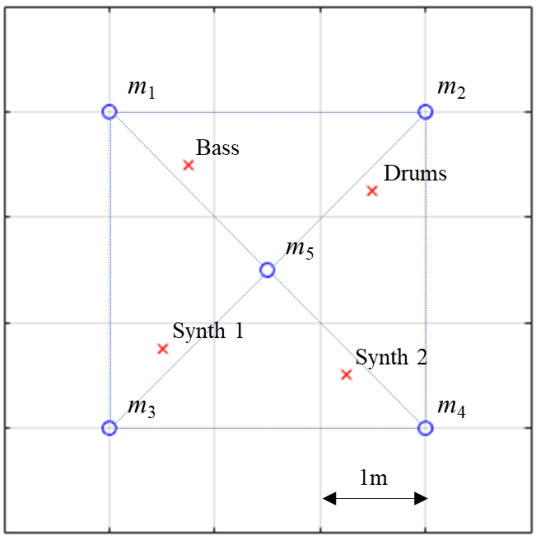


Fig. 7: Test scene (*'5 HOA'*) with 5 HOA sources. All sound sources are located inside the microphone area.

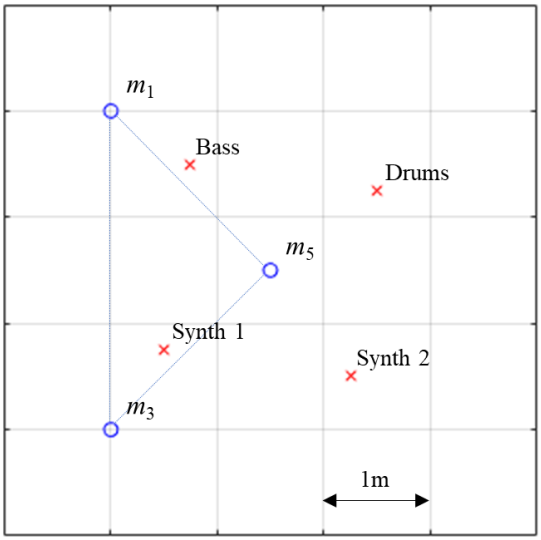


Fig. 8: Test scene (*'3 HOA'*) with 3 HOA sources, Three sound sources are located outside the microphone area.

For both scenes, the listeners were asked to rate two test items, a renderer performing exterior rendering as proposed in Section 3, labeled '*ext*', and a 6DoF renderer which positioned the listener to the closest point on the edge of the microphone area when the listener moved outside of the area, labeled '*no-ext*'. For motion inside the microphone area, the two renderers provided identical output.

A total of 8 listeners with normal hearing took part in the test. All listeners were experts in spatial audio and had had previous experience in 6DoF VR/AR subjective evaluation. All listeners were male, aged between 30 to 50 years old. The listeners were asked to rate the test items based on overall quality and focus especially on localization of the sound sources. The listeners were encouraged to explore the scenes widely, so that they would listen to the test items both inside and outside of the microphone area.

Figure 9 shows the results of the evaluation. The figure shows mean and 95% confidence intervals for both test items over both of the scenes. Overall scores are shown as well. The results indicate that the listeners preferred the renderer utilizing the proposed exterior rendering method. Listeners appreciated the improved sound source localization when outside of the microphone area. Although both renderers used the same rendering method for interior rendering, the absolute MuSCR scores for the *no-ext* renderer were quite a bit lower than for the *ext* renderer. This is perhaps explained by feedback received from the test subjects, where they mentioned that they gave more weight to the situations during the test whenever an audible difference between the renderers was heard (i.e. when the test subject was outside of the microphone area).

5 Conclusions

This work presents an extension to a previously presented method for 6DoF binaural rendering from multi-point first- or higher-order Ambisonic (FOA/HOA) recordings for exterior rendering, i.e. for the case when the listener is outside of the area covered by the microphones. A method where interpolated spatial parameters are modified based on a mapping of sound sources onto a sphere was proposed. A real-time 6DOF listening preference test demonstrates that the proposed method provides convincing performance through improved localization of both interior and exterior sound sources. This work is related to the upcoming MPEG-I immersive audio standard (ISO/IEC 23090-4).

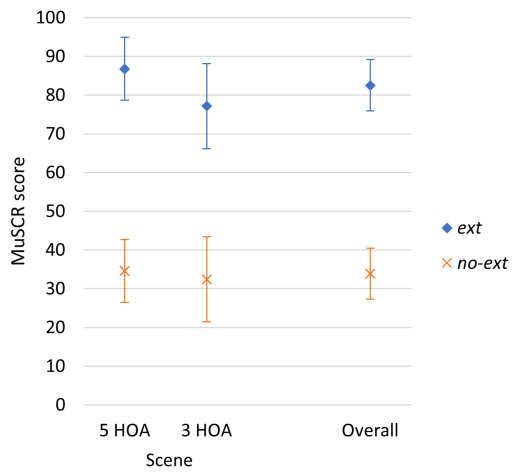


Fig. 9: Means and 95% confidence intervals for all test scenes and test items as well as averaged over test scenes.

References

[1] Politis, A., Pajunen, L., Leppanen, J., Mate, S., and Eronen, A., "Wide-Area 6DOF Rendering of Multi-Point Ambisonic Recordings Based on Interpolation of Spatial Parameters," in *Proceedings of the 2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA 2023*, IEEE, 2023.

[2] *Text of ISO/IEC CD 23090-4, MPEG-I immersive audio*, 2024, `https://www.mpeg.org/wp-content/uploads/mpeg_meetings/146_Rennes/w23717.zip`.

[3] *3GPP TS 26.253 V18.0.0 (2024-03) Codec for Immersive Voice and Audio Services - Detailed Algorithmic Description incl. RTP payload format and SDP parameter definitions (Release 18)*, 2024.

[4] McCormack, L., Politis, A., McKenzie, T., Hold, C., and Pulkki, V., "Object-based six-degrees-of-freedom rendering of sound scenes captured with multiple Ambisonic receivers," *J. Audio Eng. Soc.*, 70(5), pp. 355–372, 2022.

[5] Mariette, N. and Katz, B., "SoundDelta–large scale, multi-user audio augmented reality," in *EAA Symp. on Auralization*, volume 1, pp. 37–42, 2009.

[6] Patricio, E., Ruminski, A., Kuklasinski, A., Januszkiewicz, L., and Zernicki, T., "Toward six degrees of freedom audio recording and playback using multiple ambisonics sound fields," in *Audio Eng. Soc. Convention 146*, 2019.

[7] Schörkhuber, C., Höldrich, R., and Zotter, F., "Triplet-based variable-perspective (6DoF) audio rendering from simultaneous surround recordings taken at multiple perspectives," *Fortschritte der Akustik (DAGA)*, 4, 2020.

[8] Schultz, F. and Spors, S., "Data-based binaural synthesis including rotational and translatory head-movements," in *52nd Int. Conf. of AES*, 2013.

[9] Fernandez-Grande, E., "Sound field reconstruction using a spherical microphone array," *J. Acoustical Soc. America*, 139(3), pp. 1168–1178, 2016.

[10] Ueno, N., Koyama, S., and Saruwatari, H., "Sound field recording using distributed microphones based on harmonic analysis of infinite order," *IEEE Sig. Proc. Letters*, 25(1), pp. 135–139, 2017.

[11] Tylka, J. G. and Choueiri, E. Y., "Fundamentals of a parametric method for virtual navigation within an array of ambisonics microphones," *J. Audio Eng. Soc.*, 68(3), pp. 120–137, 2020.

[12] Birnie, L., Abhayapala, T., Tourbabin, V., and Samarasinghe, P., "Mixed source sound field translation for virtual binaural application with perceptual validation," *IEEE/ACM Trans. Audio, Speech, and Language Proc.*, 29, pp. 1188–1203, 2021.

[13] Thiergart, O., Del Galdo, G., Taseska, M., and Habets, E. A., "Geometry-based spatial sound acquisition using distributed microphone arrays," *IEEE Trans. Audio, Speech, and Language Proc.*, 21(12), pp. 2583–2594, 2013.

[14] Pezzoli, M., Borra, F., Antonacci, F., Tubaro, S., and Sarti, A., "A parametric approach to virtual miking for sources of arbitrary directivity," *IEEE/ACM Trans. Audio, Speech, and Language Proc.*, 28, pp. 2333–2348, 2020.

[15] Blochberger, M. and Zotter, F., "Particle-filter tracking of sounds for frequency-independent 3D audio rendering from distributed B-format recordings," *Acta Acustica*, 5, p. 20, 2021.

[16] *MPEG-I Immersive Audio Call for Proposals*, 2021, `https://www.mpeg.org/wp-content/uploads/mpeg_meetings/134_OnLine/w20449.zip`.

[17] Politis, A., McCormack, L., and Pulkki, V., "Enhancement of ambisonic binaural reproduction using directional audio coding with optimal adaptive mixing," in *IEEE Work. on Applications of Sig. Proc. to Audio and Acoustics (WASPAA)*, pp. 379–383, 2017.

[18] Vilkamo, J. and Pulkki, V., "Minimization of decorrelator artifacts in directional audio coding by covariance domain rendering," *J. Audio Eng. Soc.*, 61(9), pp. 637–646, 2013.

[19] McCormack, L., Politis, A., and Pulkki, V., "Rendering of source spread for arbitrary playback setups based on spatial covariance matching," in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 371–375, 2021.

[20] Zotter, F. and Frank, M., *Ambisonics: A practical 3D audio theory for recording, studio production, sound reinforcement, and virtual reality*, Springer Nature, 2019.

[21] Ivanic, J. and Ruedenberg, K., "Rotation matrices for real spherical harmonics. Direct determination by recursion," *J. Phys. Chem.*, 100(15), pp. 6342–6347, 1996.

[22] Gamper, H., "Head-related transfer function interpolation in azimuth, elevation, and distance," *J. Acoustical Soc. America*, 134(6), pp. EL547–EL553, 2013.

[23] Sporer, T., Jelfs, S., and Herre, J., "Quality Testing for AR and VR in MPEG-I Immersive Audio," *Fortschritte der Akustik (DAGA)*, 2023.