Signal-independent approach to variable-perspective (6DoF) audio rendering from simultaneous surround recordings taken at multiple perspectives

Franz Zotter[1], Matthias Frank[1], Christian Schörkhuber[2], Robert Höldrich[1],

[1] *Institute of Electronic Music and Acoustics, University of Music and Performing Arts Graz, Austria, Email: zotter@iem.at*

[2] *sonible GmbH, Graz, Austria, Email: chistian.schoerkhuber@sonible.com*

Introduction

Six-Degrees-of-Freedom audio recording and rendering approaches have been recently proposed to enable a variable-perspective playback for a listener. There are works presenting spatially sampled BRIR sets [1, 2] to prepare for the variable-position and variable-orientation binaural rendering based on either linearly interpolated binaural (dummy-head) signals [3, 4] or parametrically interpolated ones [5]. Moreover, some works present spatially distributed measurements of perspective room impulse responses [6] and we find works about projecting directionally localized sound objects in single-perspective recordings onto an outer convex hull of the room [7, 8, 9, 10], and works and patents about the interpolation from perspective recordings synchronously taken at multiple perspectives in the room, with parametric concepts to extract and render the sources detected therein and the diffuse or unlocalized parts [11, 12, 13, 14, 15, 16, 17, 18]. Some of the works avoid or at least partly avoid any short-term time-frequency-filtering based processing to get artifact-free baseline rendering [19, 20, 21, 22, 23, 16, 24], which, however, may stay limited in spatial precision.

Anyway, any kind of broadband baseline rendering method is valuable to either conceal annoying time-frequency-based processing artifacts or just as a standalone solution of decent audio quality. This contribution presents a simple and signal-independent strategy that was outlined in [19, 20] and tested in [21], but has never been written up in its most simplistic form that proved useful in practical demonstrations. It is based on recordings with distributed Oktava MK-4012 4D A-Format microphone arrays. Our demos typically used 12. . . 16 tetrahedral microphones covering a walkable area of 25. . . 200 m^2. In the virtual space, for each array, the 4 array signals are routed to 4 virtual loudspeaker objects (VLOs), and an additional objects to get a good extrapolation when the listener is located between wall and the area covered with the distributed arrays. For this particular setup, the processing steps of the rendering method are outlined in greater detail below.

Virtual Loudspeaker Objects

For each recording perspective, playback is accomplished by several loudspeaker objects of a virtual surround setup whose center is collocated with the physical recording perspective, cf. Fig. 1. Each of the loudspeaker signals is rendered by directional placement relative to the listener and by weighting with 1/r of the relative distance.

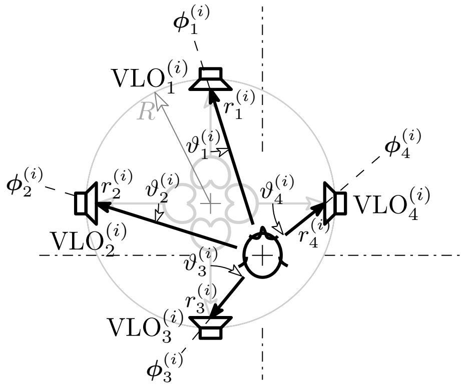

Figure 1: Virtual loudspeaker objects (VLOs) for extrapolation from a single recording perspective. From their reference radius R as a parameter, displacements are mapped to new angles $\varphi_l^{(i)}$ and gains as a function of $R/r_l^{(i)}$ and $\vartheta_l^{(i)}$.

To avoid poor imaging, additional gains are applied on every virtual loudspeaker signal to suppress its signal whenever the listener either gets too close, or stands behind the virtual loudspeaker. To avoid boundary effects when rendering to a listener outside of the area covered by the distributed perspective recordings, consistent additional surround perspectives are simulated by the image source method, simulating virtual room walls.

Single-perspective extrapolation

As in [25], microphone pairs of the Oktava 4D Ambient microphone can be used to render ORTF-like stereophonic playback. Despite the aiming of the tetrahedral microphone array elements is actually not purely horizontal, our simplistic approach maps its 4 signals purely horizontally (2D) to the azimuth $\frac{\pi}{2}l$ for a listener centered at the recording perspective $\boldsymbol{r}_i$, cf. Fig. 1. In this listening position $\boldsymbol{s} = \boldsymbol{r}_i$, each of the locally recorded signals can be represented as virtual loudspeaker direction

$$\boldsymbol{\theta}_l = \begin{bmatrix}\cos(\frac{\pi}{2}l)\\ \sin(\frac{\pi}{2}l)\end{bmatrix}. \tag{1}$$

However, rather than presenting directional signals observed at $\boldsymbol{r}_i$ to the listener, a presentation of virtual loudspeaker objects (VLOs) located at a finite distance R with regard to $\boldsymbol{r}_i$,

$$\boldsymbol{r}_l^{(i)} = \boldsymbol{r}_i + \mathrm{R}\,\boldsymbol{\theta}_l, \tag{2}$$

is more helpful, as it easily permits small parallactic shifts of the listener at $\boldsymbol{s} \neq \boldsymbol{r}_i$.

Given the position $\boldsymbol{r}_l^{(i)}$ of the l^{th} virtual loudspeaker object $\mathrm{VLO}_l^{(i)}$ of the i^{th} recording perspective in 2D, the signal of the $\mathrm{VLO}_l^{(i)}$ is presented from the distance $r_l^{(i)}$ and the direction $\boldsymbol{\phi}_l^{(i)}$

$$r_l^{(i)} = \|\boldsymbol{r}_l^{(i)} - \boldsymbol{s}\|, \qquad \boldsymbol{\phi}_l^{(i)} = \frac{\boldsymbol{r}_l^{(i)}-\boldsymbol{s}}{\|\boldsymbol{r}_l^{(i)}-\boldsymbol{s}\|} \tag{3}$$

to the listener at $\boldsymbol{s}$.

Now the question is how to present the distance and direction of each $\mathrm{VLO}_l^{(i)}$. In this contribution, we choose horizontal Ambisonics of the 3rd order to represent the direction $\boldsymbol{\phi}_l^{(i)}$ and a $\frac{1}{r}$ attenuation function to represent the distance $r_l^{(i)}$; an additional acoustic delay to represent $r_l^{(i)}$ was omitted to avoid Doppler shifts or excessive localization shifts. Still, issues arise whenever the listener either moves too close to a VLO, causing excessive gains, or too far behind a VLO, causing invalid directional mapping, cf. Fig. 2. The particular gain function outlined below circumvents these issues.

Distance and direction-dependent VLO gains

Each of the 4 VLOs per recording perspective i is supplied by one of the 4 corresponding array signals $x_1^{(i)}(t) \dots x_4^{(i)}(t)$. The virtual 4-channel layout is set up at a radius $R = 1.5\,\mathrm{m}$ with regard to the recording perspective, and the gain of each VLO is a function of the relative distance ${r_l^{(i)}}/{R}$ with regard to R, which is unity for a centered listener and smaller for a listener further away. To avoid gain increase above unity, the distance-dependent part of the gain function becomes

$$g_l^{(i)}(r) = \begin{cases} \frac{R}{r_l^{(i)}}, & \text{for } r_l^{(i)} > R \\ \frac{r_l^{(i)}}{R}, & \text{for } r_l^{(i)} \leq R. \end{cases} \tag{4}$$

This distance-dependent part is multiplied with an angular attenuation function depending on the aiming offset between the radiation direction of the $\mathrm{VLO}_l^{(i)}$ and the listener. This angular attenuation is defined as

$$\Gamma_l^{(i)}(\vartheta_l^{(i)}, r_l^{(i)}) = (1-\tfrac{\alpha}{2}) + \tfrac{\alpha}{2}\cos\vartheta_l^{(i)}, \tag{5}$$

$$\alpha = \frac{r_l^{(i)}}{r_l^{(i)}+R_{\mathrm{dir}}}, \qquad \cos\vartheta_l^{(i)} = \boldsymbol{\theta}_l^{\mathrm{T}}\boldsymbol{\phi}_l^{(i)},$$

and the distance $R_{\mathrm{dir}} = 1.1\,\mathrm{m}$ marks the transition from an omninirectional to a cardioid VLO directivity. This ensures that a listener approaching any VLO from behind receives its signal attenuated, and yet, when axially passing through a VLO, the gain won't be discontinuous. The total gain for every VLO is then

$$a_l^{(i)} = g_l^{(i)}\ \Gamma_l^{(i)}. \tag{6}$$

Directional encoding of VLO signals with gains

With the signals $x_{1\dots4}^{(i)}(t)$ from all 4 VLOs of the i^{th} recording perspective, the Ambisonic signal for the listener is obtained by multiplication with the Ambisonic encoder [26] for every VLO direction $\boldsymbol{y}_{\mathrm{N}}(\boldsymbol{\phi}_l^{(i)})$ and the signal gains $a_l^{(i)}$ as specified above, cf. Fig. 4. We obtain the corresponding $2\mathrm{N}+1$ signals for horizontal-only Ambisonics,

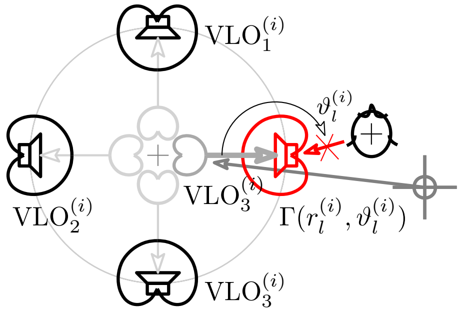

Figure 2: Virtual loudspeaker object is attenuated towards its back to avoid severe mislocalization and create a directionally consistent hull of enveloping sounds; circle/cross illustrate the mapping of an exemplary source.

$$\boldsymbol{\chi}_{\mathrm{N}}^{(i)}(t) = \sum_{l=1}^{4} a_l^{(i)}\ \boldsymbol{y}_{\mathrm{N}}(\boldsymbol{\phi}_l^{(i)})\ x_l^{(i)}(t); \tag{7}$$

while the order $\mathrm{N} = 3$ is enough for headphone playback, we used $\mathrm{N} = 5$ to keep loudspeaker playback an option.

Interpolation between Perspectives

With signals from all D recorded perspectives combined, the Ambisonic signal for the listener simply becomes

$$\boldsymbol{\chi}_{\mathrm{N}}(t) = \sum_{i=1}^{D} \boldsymbol{\chi}_{\mathrm{N}}^{(i)}(t), \tag{8}$$

As all the necessary information such as levels and directions are already contained in the single-perspective signals. While the single perspectives alone are neither precise nor perfect, their superposition will naturally contain emphasized contributions of recording perspectives closer to the listener and recorded sources, by their advantage in signal gain. However, at the outer border or outside the distributed recording perspectives, listeners would only get sounds from a half space. While this is neither naturally enveloping nor realistic, the insertion of image perspectives proves useful.

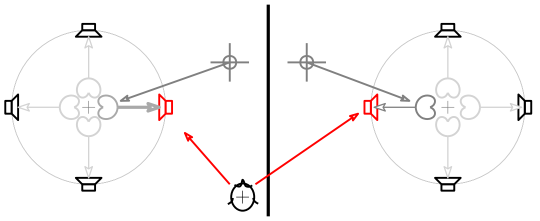

Figure 3: VLOs are mirrored for enveloping and realistic rendering outside the area of distributed perspectives; circle/cross illustrate how this maps an exemplary source.

Image recording perspectives

In our practical examples, we had 12-16 tetrahedral microphones covering a walkable area of $25\dots200\,\mathrm{m}^2$. Despite the effort, at the boundaries of this area, image recording perspectives had to be introduced to prevent a one-sided envelopment there. Fig. 3 shows an example for the images for a single recording perspective a the right boundary. Note that directional encoding and gains and of the image VLOs need to be controlled independently and the signal layout is consistently flipped in space.

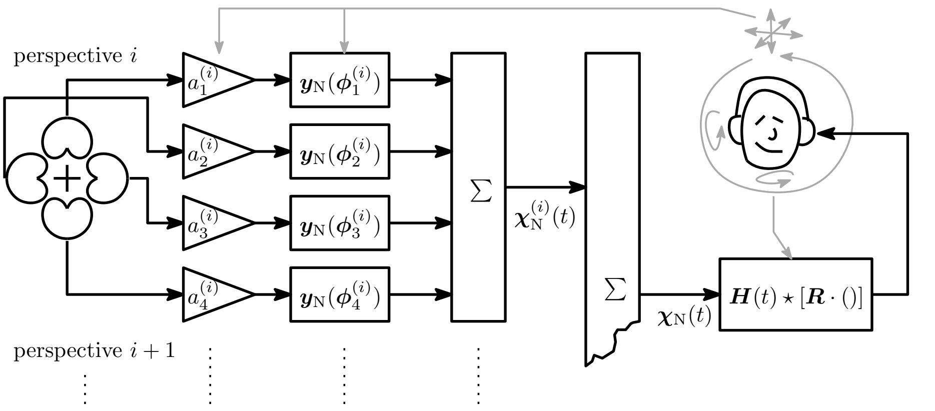
perspective i
$a_1^{(i)}$
$a_2^{(i)}$
$a_3^{(i)}$
$a_4^{(i)}$
$\boldsymbol{y}_\mathrm{N}(\boldsymbol{\phi}_1^{(i)})$
$\boldsymbol{y}_\mathrm{N}(\boldsymbol{\phi}_2^{(i)})$
$\boldsymbol{y}_\mathrm{N}(\boldsymbol{\phi}_3^{(i)})$
$\boldsymbol{y}_\mathrm{N}(\boldsymbol{\phi}_4^{(i)})$
Σ
$\boldsymbol{\chi}_\mathrm{N}^{(i)}(t)$
Σ
$\boldsymbol{\chi}_\mathrm{N}(t)$
$\boldsymbol{H}(t) \star [\boldsymbol{R} \cdot ()]$
perspective $i+1$

Figure 4: Signal chain of the proposed rendering approach.

Dynamic binaural rendering

To render variable-orientation headphone signals from the Ambisonic signal set $\boldsymbol{\chi}_\mathrm{N}$, the signals are first rotated by sample-wise multiplication with an $(2\mathrm{N}+1) \times (2\mathrm{N}+1)$ 2D Ambisonic rotation matrix $\boldsymbol{R}(\rho)$ covering the corresponding real-time captured head-tracking angle around the z axis, cf. [26]. The resulting Ambisonic signals are then matrix-convolved with a $2 \times (2\mathrm{N}+1)$ MagLS [27, 26] binaural decoder $\boldsymbol{H}(t)$

$$\boldsymbol{x}(t) = \boldsymbol{H}(t) \star [\boldsymbol{R}(\rho)\, \boldsymbol{\chi}_\mathrm{N}(t)]. \qquad (9)$$

The block diagram in Fig. 4 shows all the signal processing steps that utilize the position-dependent gains $a_l^{(l)}$, directions $\boldsymbol{\phi}_l^{(i)}$ and orientation $\boldsymbol{R}(\rho)$ found by tracking the virtual listener.

Discussion

The proposed VLO-approach basically re-encodes all the recorded signals at high spatial resolution. This choice has been made, because results can be disappointing when superimposing just a few concentric recording perspectives with only limited spatial resolution, e.g. with linear mixing of two perspectives [28] or square/triangular interpolation [24, 21]. Not only may such an approach yield relatively clear comb filter artifacts by interference of a few, non-diverse signals, it may also yield audible discontinuities when the listener is passing through lines/triangles/quadrilaterals and hereby one perspectives gets exchanged by another.

The proposed approach is also built on recorded perspectives of limited spatial resolution, however with two main differences: Their mixture only uses an distance-depending attenuation that never completely shuts off a recording perspective. Secondly, the virtual loudspeaker objects belonging to one perspective are never fixed directionally, but they parallactically shift with regard to the listener and they are weighted individually. While this leads to a somewhat enriched spatial diversity and density, this still turns out to be a way of presenting a consistent audio scene.

Nevertheless, there is potential for poor or misleading spatial mapping of sound objects in the audio scene, in particular regarding their direct sound. This is because the direction-dependent weight of the virtual loudspeaker objects (VLOs) is neither perfect nor can it re-locate the virtual sound objects. However, it largely prevents mislocalization of distant VLOs pointing away from the listener. In this way, the enveloping sounds will be consistently mapped.

While the presented method might seem highly simplistic, the basic performance has been tested in [21] and has outperformed a signal-independent triangular interpolation method.

Conclusion

In this contribution, we presented a simple baseline implementation of our practice-proof user-navigable renderer according to [19] that is robustly providing position-dependent and orientation-dependent rendering of the highest audio quality. The spatial mapping of the algorithm is simplistic, signal-independent (non-parametric), and except the HRIR part for binaural rendering, all its signal processing is frequency-independent.

While the approach may not always provide the highest-possible spatial fidelity, it has still been rated better than other simplistic approaches such as a signal-independent triangular interpolation [21].

By design, its avoidance of parametric time-frequency processing always robustly preserves a high audio quality. Our implementation even omits the option of time-delays mentioned in our patent [19] in order to avoid Doppler shifts or time-variant interference for a moving listener.

We believe that the algorithm is practical to be used as is, and moreover it is perfectly suited for combination with signal-dependent approaches. In such more complex, time-frequency-dependent, parametric rendering scenarios, our signal-independent renderer offers to be an error-concealing complement when mixed with, or it can be used to provide rendering of residuals for which no parametric mapping could be found.

Extensions to more elaborated microphone arrays at the recording perspectives, such as, e.g. Zylia ZM-1 [22] or ESMA [29], to 3D navigation, etc. are straightforward from the given descriptions.

References

[1] A. Neidhardt, "Data set: Brirs for position-dynamic binaural synthesis measured in two rooms," in *Proc. ICSA 2019*, 2019.

[2] B. Bacila and H. Lee, "360 degree binaural room impulse response (brir) database for 6dof spatial perception research," in *e-Brief 513 AES Conv*, Dublin, 2019.

[3] A. Neidhardt and N. Knoop, "Binaural walk-through scenarios with actual self-walking using an htc vive," in *Fortschritte der Akustik - DAGA*, Kiel, 2017.

[4] S. Werner, F. Klein, and G. Götz, "Investigation on spatial auditory perception using non-uniformspatial distribution of binaural room impulse responses," in *Proc. ICSA 2019*, 2019.

[5] V. Garcia-Gomez and J. J. Lopez, "Binaural room impulse responses interpolation for multimedia real-time applications," in *prepr. 9962 in AES Conv.*, Milan, 2018.

[6] R. Stewart and M. Sandler, "Database of omnidirectional and b-format room impulse responses," in *Proc. ICASSP*, Dallas, June 2010.

[7] V. Pihlajamäki, Tapani; Pulkki, "Synthesis of complex sound scenes with transformation of recorded spatial sound in virtual reality," *JAES*, vol. 7/8, no. 63, pp. 542–551, 2015.

[8] A. Plinge, S. J. Schlecht, O. Thiergart, T. Robotham, O. Rummukainen, and E. A. P. Habets, "Six-degrees-of-freedom binaural audio reproduction of first-order ambisonics with distance information," in *AES Int. Conf. Audio f. Virt. and Aug. Reality*, 2018.

[9] A. Allen and B. Kleijn, "Ambisonic soundfield navigation using directional decomposition and path distance estimation," in *Proc. ICSA*, Graz, 2017.

[10] A. Allen, "Ambisonic sound field navigation using directional decomposition and path distance estimation," *US Patent*, no. US 10,182,303 B1, 2019.

[11] G. D. Galdo, O. Thiergart, T. Weller, and E. Habets, "Generating virtual microphone signals using geometrical information gathered by distributed arrays," in *Proc. IEEE Workshop HSCMA*, Edinburgh, 2011.

[12] O. Thiergart, G. D. Galdo, M. Taseska, and E. Habets, "Geometry-based spatial sound acquisition using distributed microphone arrays," *IEEE TASLP*, vol. 21, no. 12, 2013.

[13] J. G. Tylka and E. Y. Choueiri, "Comparison of techniques for binaural navigation of higher-order ambisonic soundfields," in *prepr. 9421 AES Conv.*, New York, 2015.

[14] ——, "Soundfield navigation using an array of higher-order ambisonics microphones," in *AES Int. Conf. AVAR*, Los Angeles, 2016.

[15] ——, "Models for evaluating navigational techniques for higher-order ambisonics," in *Proc. ASA Meeting*, Boston, 2017.

[16] ——, "Domains of practical applicability for parametric interpolation methods of virtual sound field navigation," *JAES*, vol. 67, no. 11, pp. 882–893, 2019.

[17] ——, "Performance of linear extrapolation methods for virtual sound field navigation," *JAES*, vol. 68, no. 3, 2020.

[18] ——, "Fundamentals of a parametric method for sound field navigation within an array of ambisonics microphones," *JAES*, vol. 68, no. 3, pp. 120–137, 2020.

[19] P. Grosche, F. Zotter, C. Schörkhuber, M. Frank, and R. Höldrich, "Method and apparatus for acoustic scene playback," *WO Patent*, no. WO 2018/077379 A1, 2018.

[20] T. Deppisch and A. Sontacchi, "Browser application for virtual audio walkthrough," in *Forum Media Technology & All Around Audio*, St. Pölten, 2017.

[21] D. Rudrich, M. Frank, and F. Zotter, "Evaluation of interactive localization in virtual acoustic scenes," in *Fortschritte der Akustik - DAGA*, Kiel, 2017.

[22] E. Patricio, A. Ruminński, A. Kuklasiński, Ł. Januszkiewicz, and T. Żernicki, "Toward six degrees of freedom audio recording and playback using multiple ambisonics sound fields," in *prepr. 10141 AES Conv*, Dublin, 2019.

[23] D. R. Méndez, C. Armstrong, J. Stubbs, M. Stiles, and G. Kearney, "Practical recording techniques for music production with six-degrees of freedom virtual reality," in *prepr. 464 AES Conv.*, New York, 2018.

[24] N. Mariette and B. F. Katz, "Sounddelta – large scale, multi-user audio augmented reality," in *EAA Symposium on Auralization*, Espoo, June 2009.

[25] E. Kurz, F. Pfahler, and M. Frank, "Comparison of first-order ambisonic microphone arrays," in *Int. Conf. Spatial Audio (ICSA)*, Graz, September 2015.

[26] F. Zotter and M. Frank, *Ambisonics.* SpringerOpen, 2019. [Online]. Available: https://doi.org/10.1007/978-3-030-17207-7

[27] C. Schörkhuber, M. Zaunschirm, and R. Höldrich, "Binaural Rendering of Ambisonic Signals via Magnitude Least Squares," in *Fortschritte der Akustik - DAGA*, Munich, March 2018.

[28] A. Southern, J. Wells, and D. Murphy, "Rendering walk-through auralisations using wave-based acoustical models," in *EUSIPCO*, Glasgow, August 2009.

[29] H. Lee, "Capturing 360° audio using an equal segment microphone array (esma)," *JAES*, vol. 67, no. 1/2, pp. 13–26, 2019.