PROCEEDINGS of the
24th International Congress on Acoustics

October 24 to 28, 2022 in Gyeongju, Korea

ABS-0617

# Parametric binaural reproduction of higher-order spatial impulse responses

Christoph Hold[(1)], Leo McCormack[(1)], Ville Pulkki[(1)]

[(1)]Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland, First.Last@aalto.fi

**ABSTRACT**

Spherical microphone arrays may be used to capture the directional characteristics of a room acoustic response. Spatial impulse response rendering (SIRR) is a method for parameterizing the response in terms of its principle directional and diffuse components, which allows for subsequent spatially enhanced reproduction of these captured spatial characteristics. This paper explores a reformulation of the higher-order spatial impulse response rendering (HO-SIRR) method for direct-to-binaural reproduction. The previously proposed HO-SIRR formulation was optimized for loudspeaker-based reproduction. While the loudspeaker channels may indeed be binauralized for headphones playback, such an approach does not necessarily take advantage of the full resolution of the employed head-related transfer function (HRTF) set and may incur coloration due to the employed amplitude panning functions. Therefore, this paper proposes a reformulation of the HO-SIRR for binaural rendering, which addresses these concerns. The proposed method is evaluated through objective perceptual metrics, which highlight reduced coloration error, while maintaining minimal spatial error only imposed by the HRTF measurement grid. An open-source implementation of the proposed reformulation is also made available.

Keywords: Spatial Impulse Response, Higher-Order Ambisonics, Spatial Audio, Binaural Rendering

## 1 INTRODUCTION

Capturing and reproducing the spatial properties of acoustical spaces finds a number of different applications, including the perceptual evaluation of concert halls and other spaces of interest [5], and for example for artistic productions [1]. Spatial room impulse response (RIR) rendering algorithms facilitate this need, by rendering RIRs corresponding to a target loudspeaker or headphones setup, based on a microphone array RIR as input. After convolving a monophonic signal with the rendered RIR, the input signal is then reproduced over the target playback setup, while also exhibiting much of the spatial characteristics of the captured space. The spatial resolution of this rendering is dictated by the spatial RIR rendering method and hence is an active area of research.

There are a number of linear signal-independent methods that are applicable for this task. One popular approach is to first convert the input microphone array RIR into the spherical harmonic domain [10], which is referred to as encoding in Ambisonics terminology. These spherical harmonic (or Ambisonic) RIRs may then be rendered to the target playback setup by using one of a number of available Ambisonic decoding approaches [11]. However, given that many available microphone arrays are limited to lower-order Ambisonics recording, and that there is an upper limit to what can be achieved with purely linear and signal-independent mappings of channels, there have been a few proposals for signal-dependent alternatives to this spatial RIR rendering task, which aim to improve the spatial resolution of the rendering.

The first proposed signal-dependent method for spatial RIRs, which is often also referred to as a parametric method, was the Spatial Impulse Response Rendering (SIRR) method [8]. The method is based on a parametric sound-field model, which assumes a mixture of a single reflection and isotropic diffuse reverberation per time-frequency tile. The method operates based on active-intensity based analysis techniques, which are used to estimate the direction-of-arrival (DoA) of the single reflection and also a diffuseness estimate. The omni-directional component is then panned to the target loudspeaker setup using VBAP, and decorrelated versions are also routed to all loudspeaker channels. The diffuseness estimates are then used to balance between these direct

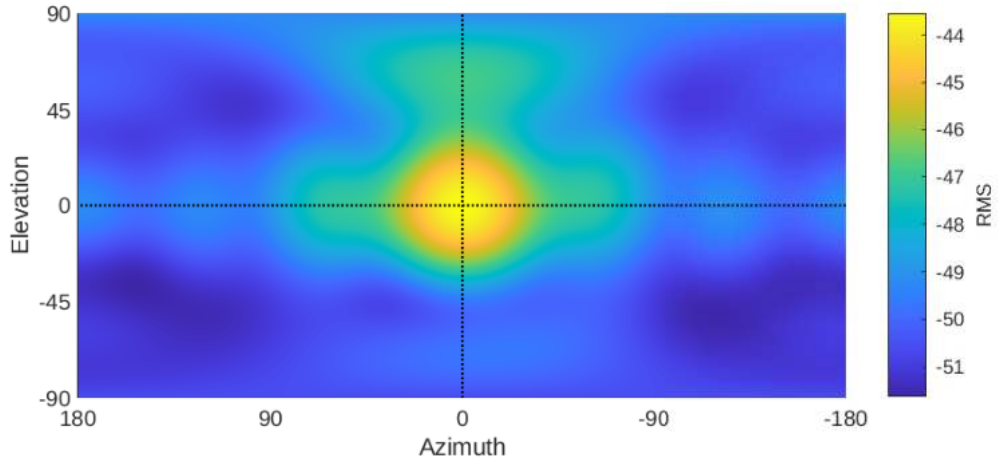


Figure 1. Input RMS in dB, evaluated on a dense grid, from a third order Spatial Impulse Response (SIR).

and diffuse streams.

Recently, the SIRR method was extended to support higher-order Ambisonic input (HO-SIRR) [7] by dividing the sound-field into directionally-constrained sectors, in the same manner as its signal-domain counterpart: higher-order Directional Audio Coding [9]. The DoA and diffuseness estimates are then made for each sector, and the sector signals are panned to the target loudspeaker setup accordingly for the direct stream rendering. For the diffuse stream, HO-SIRR instead scales the sector components by the diffuseness estimates, re-encodes them into the Ambisonics format, reproducing it using an Ambisonic decoder, followed by decorrelation.

A recent reformulation of higher-order Directional Audio Coding [3] uncovered several shortcomings regarding the re-encoding approach. This article is a first step towards implementing improvements and new techniques to HO-SIRR.

The current formulation of HO-SIRR described in [7] is optimized specifically for loudspeaker based rendering. While it is possible to binauralize a virtual loudspeaker setup, there are three main issues that may arise. First, a viable virtual loudspeaker grid is only of finite resolution, while the HRTFs are usually available for a much denser grid or even a spatially continuous resolution. Second, the inherent amplitude panning in loudspeaker based rendering can introduce coloration, especially on an arrangement typical for binaural reproduction. Last, the additional complexity can be avoided by utilizing the target rendering format, binaural output, directly.

## 2 PROPOSED METHOD

The proposed method is divided into dedicated analysis and synthesis stages. During the analysis stage, a set of spatial parameters are extracted from the input spatial RIR. An example of a spatial RIR is provided in Fig. 1, which depicts the directional amplitude density of the input sound-field. For the present example it shows a strong frontal peak with multiple distinct lower energy peaks to the sides and the top.

The higher-order input RIR is first transformed into the time-frequency domain and then divided into spherical sectors $\xi$ of each first order components. This method allows simultaneous extraction of multiple pseudo-intensity vectors, each estimating the principal DoA of its sector with the following:

$$\boldsymbol{i}_\xi \propto \Re\{p_\xi^{\mathsf{H}} \boldsymbol{v}_\xi\}\,. \tag{1}$$

$$\Omega_\xi^{\mathrm{DoA}} = \angle \boldsymbol{i}_\xi\,. \tag{2}$$

The diffuseness estimate $\psi_\xi$ is then obtained as

$$\psi_\xi = 1 - \frac{||\boldsymbol{i}_\xi||}{E_\xi} = 1 - \frac{||\boldsymbol{i}_\xi||}{|p_\xi|^2 + \boldsymbol{v}_\xi{}^{\mathsf{H}} \boldsymbol{v}_\xi}\,. \tag{3}$$

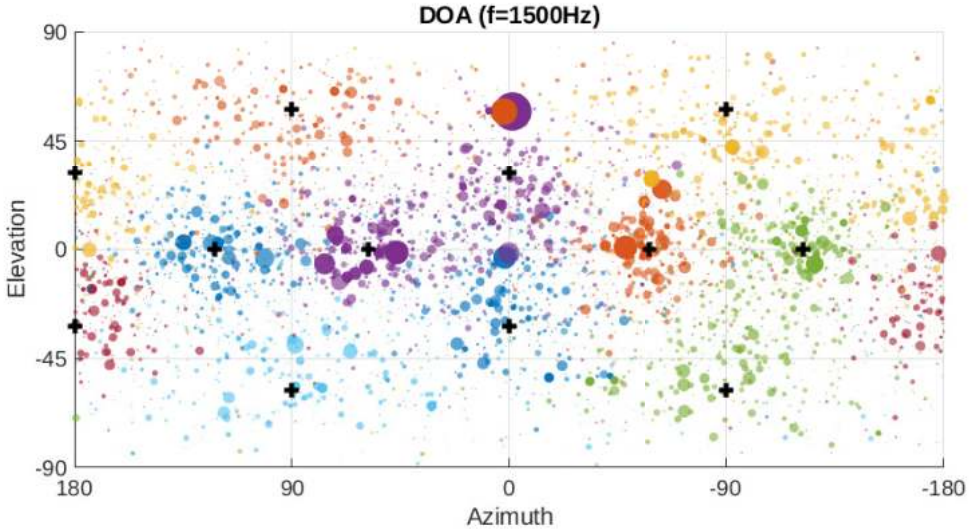


Figure 2. Visualization of HO-SIRR analysis in a single frequency band. Color corresponds to the sector, size scales with energy, and opacity inversely proportional to diffuseness. Black crosses mark sector steering directions from third order Spatial Impulse Response (SIR).

Given the input Ambisonic RIR depicted Fig. 1, an visualisation of the corresponding analysis data is provided in Fig. 2. The plot demonstrates the spatial separation between sector estimates, while still allowing soft transitions. The figure also shows that the estimates are clustered around their corresponding sector steering direction, which stems from the directional weighting introduced by the sector pattern. The size and opacity of the estimate may provide an intuitive indication regarding their relative importance, since their size correspond to the energy of each reflection, and the reflections associated with a high diffuseness value are rendered as more opaque; thus indicating a less reliable DoA estimate.

The synthesis is carried out based on spatializing the sector pressure signals $p_\xi$. Directional components are convolved with HRTFs corresponding to the estimated DoAs $\Omega_{\mathrm{DoA},\xi}$. The (locally) diffuse components are not directionally manipulated and thus rendered according to their originating sector steering direction $\Omega_\xi$. Furthermore, decorrelation may be applied to the sector pressure signals, for example through phase randomization, yielding $\tilde{p}_\xi$. The mixing between the directional and diffuse components is determined by the estimated sector diffuseness $\psi_\xi$.

The rendering may therefore be formulated, for each ear $l, r$, as

$$H(\omega)^{l,r} = \sum_{\xi=1}^{J} \beta_A \sqrt{(1-\psi_\xi)}\, p_\xi(\omega)\mathrm{HRTF}^{l,r}(\Omega_{\mathrm{DoA},\xi}) + \beta_E \sqrt{\psi_\xi}\, \tilde{p}_\xi(\omega)\mathrm{HRTF}^{l,r}(\Omega_\xi)\,. \tag{4}$$

Note that the time dependency of the STFT was omitted for brevity of notation. The preservation factor $\beta_A$ restores the scaling of the sum of beamformer outputs $p_\xi$ and factor $\beta_E$ for the decorrelated components. These preservation factors, their derivations, and application are explained in [2, 4]. Further details about the algorithm and the estimated spatial parameters may be found in [7].

The beamformers extracting $p_\xi$ are arranged in a specific way that ensures uniform coverage of the sphere, while utilizing the $\max_{rE}$ pattern to optimize directivity. The arrangement is therefore also referred to as a spherical filter bank (SFB), since it divides the spherical input sound-field, given these preservation objectives, such that all of the input sound-field information is retained within the sector signals.

This presented approach mitigated the aforementioned coloration issues, and more elaborate methods, for example based on the SFB reproduction properties, are left for future work.

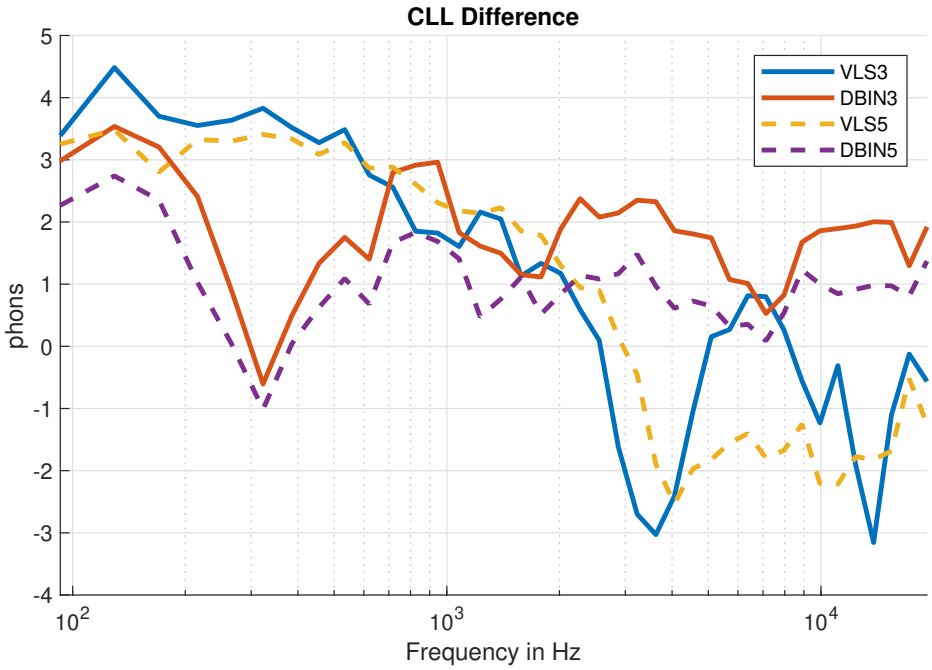


Figure 3. Composite Loudness Level difference ($\Delta CLL$) compared to a reference, evaluated in equivalent rectangular bandwidth (ERB) bands.

## 3 EVALUATION AND DISCUSSION

We consider the rendering of a high SH expansion, with state-of-the-art binaural decoding as the reference in the present scenario. The reference is set to be an impulse response encoded to eighth order and rendered by the magnitude-least-squares approach. For such high orders in combination with advanced binaural decoding, the perceptual impact of Ambisonic encoding and binaural rendering is assumed to be small [6]. HO-SIRR aims to recreate this high-order impression, which is not available in practice, by enhancing a lower order input. The rendering method can thus be seen as a spatial upmixing approach, or equivalently, as a spatial bandwidth extension. HO-SIRR achieves this by the parametrization of the input soundfield, and the interpretation that directional components can be expressed as plane-waves, which can be parameterized and recreated using arbitrarily high orders. This assumption matches selecting a single HRTF for the directional components. We can therefore conclude, that the spatial rendering error is bound by the resolution of the utilized HRTF set. Since many HRTF sets are available in very high spatial resolution, or even in the spatially continuous SH representation, we assume this error to be below the perceptual threshold and thus negligible.

The other issue we identified with rendering loudspeaker-based HO-SIRR on a virtual loudspeaker setup is related to spectral coloration. Therefore, the following presents an evaluation based on an objective perceptual metric targeting coloration. The Composite Loudness Level (CLL) provides a model of the relative perceived loudness and is comprised of the logarithm of adding both ears' radical of each RMS value. Comparing the CLL over equivalent rectangular bandwidth (ERB) bands gives a robust indicator of spectral coloration between two binaural signals. Figure 3 shows the results of $\Delta CLL = CLL_{\text{Item}} - CLL_{\text{Ref}}$ over frequency bands. The results underline the impression of the authors that the binaural formulation proposed herein shows less coloration compared to the virtual loudspeaker approach.

It is noted, however, that such model-based evaluation of perceptual dimensions is limited, and thus a formal perceptual evaluation of the reformulated method is a topic of future work. To help facilitate such future studies, an open-source implementation of the proposed reformulation is also made available in the original toolbox: https://github.com/leomccormack/HO-SIRR.

## 4 CONCLUSIONS

In this contribution a reformulation of the HO-SIRR method was explored, which achieves direct binaural rendering of higher-order Ambisonic spatial impulse responses. It is demonstrated through perceptually-motivated objective metrics that the proposed reformulation yields high perceptual quality, by incorporating recent advances in spatial audio processing. This paper hence serves as a foundation for future studies.


## ACKNOWLEDGEMENTS

Thanks are extended to Pedro Lladó and Sebastian Schlecht for the discussions that transpired during the development of the proposed reformulation.


## REFERENCES


[1] F. K. Brian, D. Poirier-Quinot, and J.-M. Lyzwa. La vierge 2020: Reconstructing a virtual concert performance through historic auralisation of notre-dame cathedral. In *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, pages 1–9, 2021.

[2] C. Hold, A. Politis, L. McCormack, and V. Pulkki. Spatial filter bank design in the spherical harmonic domain. volume 2021-Augus, pages 106–110. IEEE, 8 2021.

[3] C. Hold, V. Pulkki, A. Politis, and L. McCormack. Compression of higher-order ambisonic signals using directional audio coding. *Submitted for Review*, 0 2022.

[4] C. Hold, S. J. Schlecht, A. Politis, and V. Pulkki. Spatial filter bank in the spherical harmonic domain: Reconstruction and application. volume 2021-Octob, pages 361–365. IEEE, 10 2021.

[5] T. Lokki, J. Pätynen, A. Kuusinen, and S. Tervo. Concert hall acoustics: Repertoire, listening position, and individual taste of the listeners influence the qualitative attributes and preferences. *The Journal of the Acoustical Society of America*, 140:551–562, 2016.

[6] T. Lübeck, H. Helmholz, J. M. Arend, C. Pörschmann, and J. Ahrens. Perceptual evaluation of mitigation approaches of impairments due to spatial undersampling in binaural rendering of spherical microphone array data. *AES: Journal of the Audio Engineering Society*, 68:428–440, 6 2020.

[7] L. McCormack, V. Pulkki, A. Politis, O. Scheuregger, and M. Marschall. Higher-order spatial impulse response rendering: Investigating the perceived effects of spherical order, dedicated diffuse rendering, and frequency resolution. *Journal of the Audio Engineering Society*, 68:338–354, 6 2020.

[8] J. Merimaa and V. Pulkki. Spatial impulse response rendering i: Analysis and synthesis. *AES: Journal of the Audio Engineering Society*, 53:1115–1127, 2005.

[9] A. Politis, J. Vilkamo, and V. Pulkki. Sector-based parametric sound field reproduction in the spherical harmonic domain. *IEEE Journal of Selected Topics in Signal Processing*, 9:852–866, 8 2015.

[10] B. Rafaely. *Fundamentals of Spherical Array Processing*, volume 16. Springer International Publishing, 2019.

[11] F. Zotter and M. Frank. *Ambisonics*. Springer Topics in Signal Processing, 2019.