Parametric Reproduction of Room Impulse Responses

Valtteri Kallinen
Aalto University
Master's Programme CCIS / AAT

Valtteri.Kallinen@aalto.fi

Abstract

This paper shortly reviews the history and development of the most common parametric room impulse response (RIR) reproduction methods: SIRR, HO-SIRR and SDM. The original papers describing the techniques are considered, as well as a more recent paper examining the limitations and capabilities of SDM.

1 Introduction

Parametric room impulse response (RIR) reproduction techniques were born out of the desire to avoid the pitfalls associated with the other types of RIR reproduction methods [6]. For example, channel based RIRs captured with regular microphone arrays are difficult to generalize over different playback setups (ideally one loudspeaker per microphone), and RIR reproduction using low-order Ambisonics has problems with highly coherent loudspeaker signals causing comb filtering effects and other unwanted artifacts [1–3,11–13,16].

Thus, parametric techniques of room impulse response reproduction are intended to be applicable over different sound capture and simulation setups (regular microphones or spherical harmonics) [14]. Similarly, the reproduced response can, then, be rendered using any spatial rendering method (e.g., VBAP, Ambisonics, WFS) in the desired spatial resolution [14].

This seminar paper is organized in to sections describing the history and theory of the most popular [4] parametric methods used for RIR reproduction from the past 20 years. Section 2 discusses the spatial impulse response rendering (SIRR), Section 3 continues with the higher-order formulation of SIRR, higher-order spatial impulse response rendering (HO-SIRR), and Section 4 discusses the slightly different approach of spatial decomposition method (SDM). Section 5 concludes the paper.

2 Spatial Impulse Response Rendering

SIRR was the first method of reproducing room impulse responses parametrically using the knowledge of psychoacoustic [5]. Later, the principles of the analysis and synthesis of SIRR were applied in developing the directional audio coding (DirAC) [10].

In [6], the method is described as implemented for B-format microphones and employing STFT-based analysis. Figure 1 shows a block diagram of the analysis scheme. Depicted in the figure, the 3-D intensity vectors are calculated from the spectra in the following manner:

$$\boldsymbol{I}_a(\omega) = \frac{\sqrt{2}}{Z_0}\Re\{W^*(\omega)\boldsymbol{X}'(\omega)\} \tag{1}$$

In Equation 1, $Z_0 = \rho_0 c$ is the acoustic impedance of the medium, $W^*(\omega)$ is the complex conjugate of the Fourier transform taken from the omnidirectional signal $W(t)$. $\boldsymbol{X}'(\omega)$ is the Fourier transform of the transpose vector of the x, y and z components of the B-format signal: $X(t)$, $Y(t)$ and $Z(t)$. The energy density, shown in the bottom-right corner of Figure 1 is given by:

$$E(\omega) = \rho_0[Z_0^{-2}\ |W(\omega)|^2 + |\boldsymbol{X}'(\omega)|^2] \tag{2}$$

From the 3-D intensity vectors, the azimuth $\theta(\omega)$ and elevation $\phi(\omega)$ for the direction of arrival (DOA, opposite to the direction of $\boldsymbol{I}_a(\omega)$) can then be estimated:

$$\theta(\omega) = \tan^{-1}\left[\frac{-I_y(\omega)}{-I_x(\omega)}\right] \tag{3}$$

$$\phi(\omega) = \tan^{-1}\left[\frac{-I_z(\omega)}{\sqrt{I_x^2(\omega) + I_y^2(\omega)}}\right] \tag{4}$$

Further depicted in Figure 1, and described in [6], are the operations for calculating the diffuseness estimate $\psi(\omega)$ from the intensity vectors $\boldsymbol{I}_a(\omega)$ and energy density $E(\omega)$:

$$\psi(\omega) = 1 - \frac{\|\boldsymbol{I}_a(\omega)/c\|}{E(\omega)} \tag{5}$$

For the diffuseness estimate, $\psi = 1$ describes a completely diffuse field and $\psi = 0$ implies a sound field without oscillating energy [6].

From the analysis, the resulting parameters are then used to process the omnidirectional and to reproduce the RIR. Figure 2 describes the SIRR synthesis process, where the omnidirectional signal is split into N frequency channels for processing. As depicted next

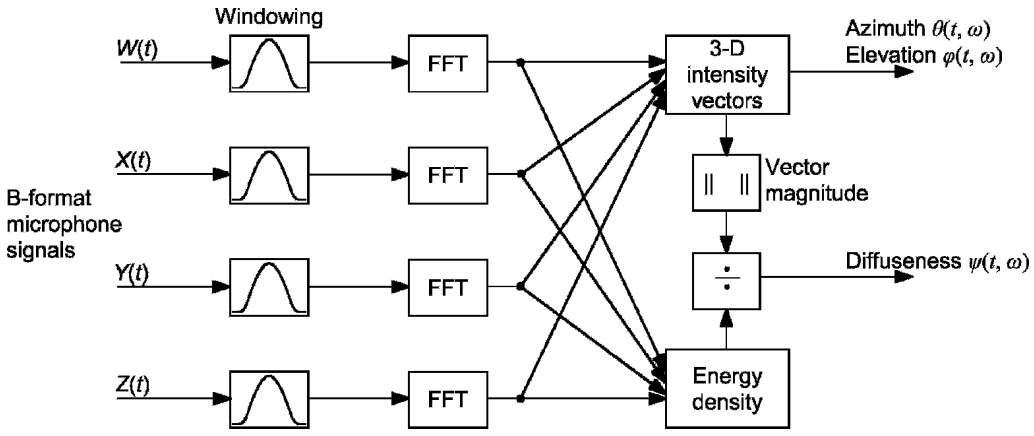

Figure 1: SIRR: analysis [6]

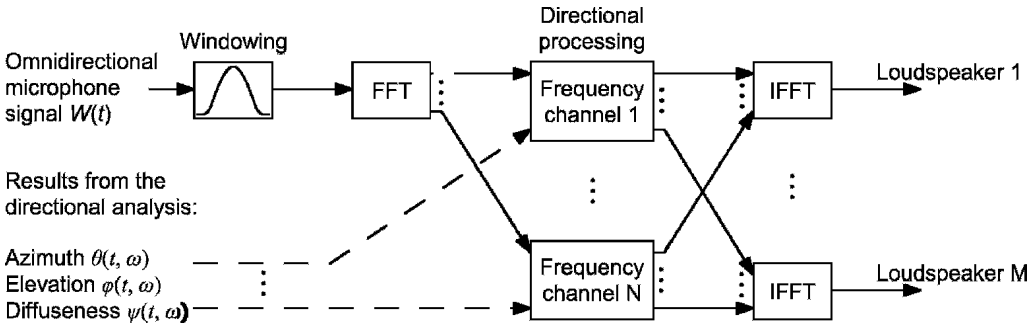

Figure 2: SIRR: synthesis [6]

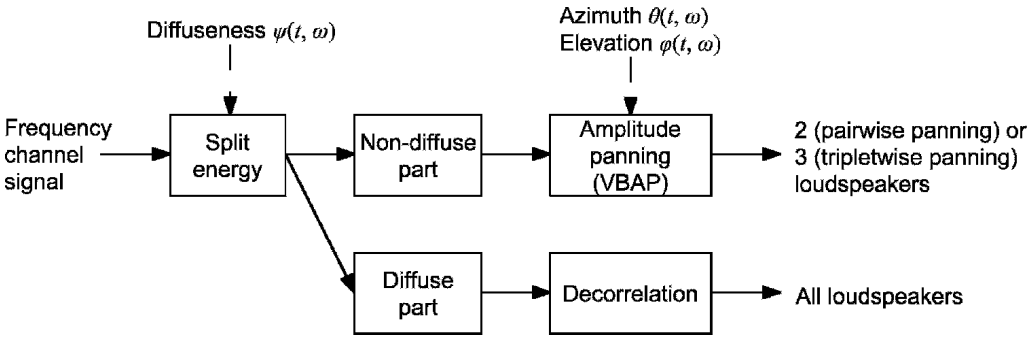

Figure 3: SIRR: synthesis, single frequency channel processing [6]

in Figure 3, each channel is then further split into non-diffuse and diffuse parts based on this frequency channel's specific diffuseness estimate from the previous analysis. The non-diffuse part of the signal is mapped to a pair or a triplet of loudspeakers (depending on the loudspeaker setup, i.e., 2D or 3D) using vector base amplitude panning (VBAP) using the DOA estimate. The diffuse part is fed to all loudspeakers and decorrelation is applied to prevent coherence and coloration issues [9].

To test the reproduction quality of SIRR, two listening tests were conducted [8]. One listening test was conducted in an anechoic chamber and another in a standard listening room. During these tests, they used a hybrid diffuse model for SIRR, producing the low frequencies using amplitude panning and the high frequencies decorrelated using phase randomization.

The anechoic listening room test compared the performance of SIRR against using only phase randomization (i.e., same as SIRR, but assuming the sound field is completely diffuse all the time) and first-order Ambisonics. The playback setup consisted of 16 Genelec 1029A loudspeakers arranged in a spherical setup as shown in Figure 4. However, the Ambisonics samples only used 4 loudspeakers in a quadraphonic setup, as it was deemed to produce better results based on previous research. In short, the outcome of the listening tests was that SIRR performed better than phase randomization alone or first-order Ambisonics.

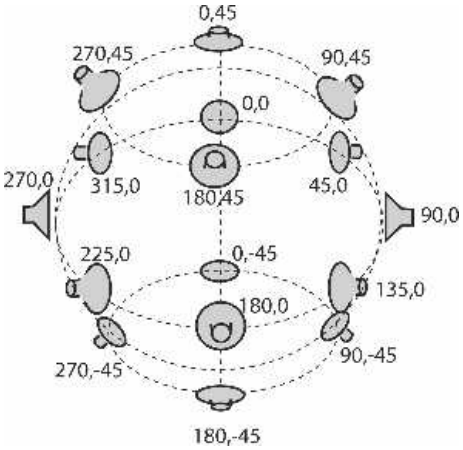

Figure 4: The loudspeaker arrangement used in the anechoic room tests [8].

The standard listening room test measured the RIR reproduction using impulse responses obtained from real acoustic environments. Since the responses were based on real measurements, there was no feasible way to compare the results with a reference. Therefore, in this test, they instead asked participants to rate the perceived 'naturalness' of the resulting reverberation. For this listening test, they used Genelec 1030A loudspeakers in a 7.0 setup depicted in Figure 5. The figure also shows the different listening positions that were used during the test. Five reproduction methods were compared: SIRR 5.0, SIRR 7.0, first-order Ambisonics, first-order Ambisonics with direct sound applied to a single loudspeaker, and omnidirectional signal to all channel with direct sound applied to a single loudspeaker. Despite having 7 loudspeakers, only the SIRR 7.0 method employed all the available channels. In the end, the results still showed better performance for SIRR compared to the other methods. However, the obtained results indicated that the reproduction quality was between 'fair' and 'good'.

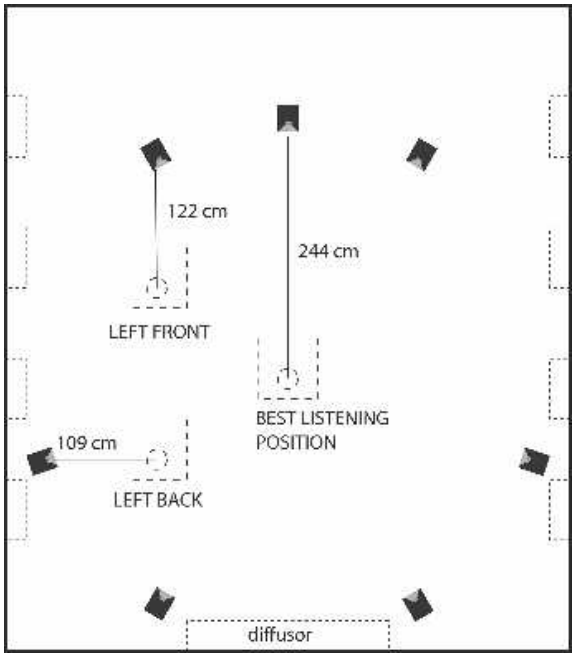

Figure 5: The loudspeaker arrangement used in the standard listening room tests [8].

3 Higher-order Spatial Impulse Response Rendering

HO-SIRR builds on the original SIRR method by considering higher order spherical harmonics domain (SHD) input in segments [4]. The segments are uniformly distributed over the sound-field. The key idea behind this segmentation was that it would improve the reproduction quality for early reflection arriving at the listener position in the same time-frequency window, solving the problem of accidentally assigning higher diffuseness values to nondiffuse sounds.

Figure 6 displays a block diagram of the HO-SIRR technique. First, the input signal is windowed in time domain and separated in to frequency components, as in the original SIRR method. However, given higher-order input, HO-SIRR splits each frequency bin to spacial sectors. Each spatial sector is then processed similarly as in the original SIRR method with the difference, compared to the original process, that the diffuse stream is encoded back to SHD. During the synthesis, the nondiffuse channels are then combined with the decoded and decorrelated diffuse stream.

That is, the direct, nondiffuse, stream is formed by panning each beamformer's first component (omnidirectional for first order) using VBAP and summing the results:

$$\mathbf{y}_{\text{dir}} = \sum_{s=1}^{S} \sqrt{\frac{1-\psi_s}{S}} \mathbf{g}(\phi_s, \theta_s) \mathrm{z}_{00}^{(s)} \tag{6}$$

Where, S is the number of sectors and ψ_s, ϕ_s and θ_s are the sector specific estimated parameters. The diffuse SHD stream is formed using:

$$\mathbf{a}_{\text{diff}} = \begin{cases} \sqrt{\frac{\psi_1}{(N+1)^2}} \mathbf{a}_1 & \text{if } N = 1 \\ \sum_{s=1}^{S} \sqrt{\frac{\psi_s}{S}} \mathbf{y}_{N-1}^{(s)} z_{00}^{(s)} & \text{if } N > 1 \end{cases} \tag{7}$$

Where, N is the spherical harmonics order and $\boldsymbol{y}_{N-1}^{(s)}$ the spherical harmonics weights corresponding to the direction of sector s. The diffuse loudspeaker signals are then formed by linearly decoding the Ambisonics signal and decorrelation:

$$\mathbf{y}_{\text{diff}} = \mathcal{D}[\mathbf{D}_{\text{ls}} \mathbf{a}_{\text{diff}}] \tag{8}$$

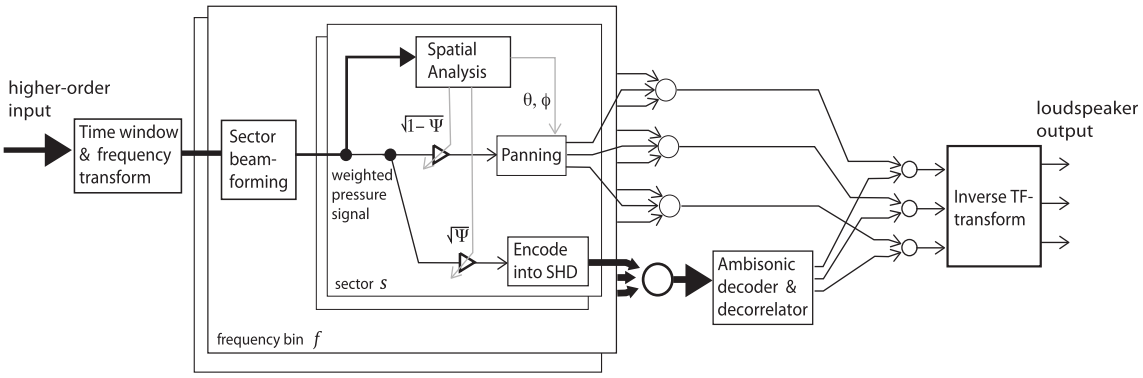

Figure 6: HO-SIRR: analysis and synthesis [4]

Where, $\mathbf{D}_{\text{ls}}$ is the sampling Ambisonics decoder matrix and $\mathcal{D}$ denotes the decorrelation. A nice property of this higher-order formulation reduces to the original SIRR method for first-order input [4].

In [4], they used a spherical array of 64 loudspeakers in an anechoic room to compare the performance of different configurations of the HO-SIRR method with B-format SDM and Ambisonics (orders 1, 3 and 5). The comparison was made using a formal listening test that was split in two parts: one to determine the effect of dedicated diffuse stream rendering and frequency-resolution, and another to test the effect of spherical harmonic input order on the perceived accuracy.

Figure 7 shows the results of the first test of [4]. The effect of dedicated diffuse stream rendering was tested by comparing SDM and the first order (legacy) SIRR with and without diffuse stream. The effect of frequency-resolution was compared by using different methods of splitting the signal in to the spectral components. The tested methods were broad-band (BB), six Equivalent Rectangular Bandwidth (6-ERB), and 128 uniformly spaced bins (128-bins). Results of the first test indicated that the effect on perceived quality from increased frequency resolution was less dramatic compared to the added dedicated diffuse stream rendering [4].

The second test results (depicted in Figure 8) showed that increased spherical harmonics order of the input signal improved the perceived accuracy of the RIR reproduction. The score improved for higher orders of SIRR, although the effect was diminished for more transient signals (kick drum). Moreover, the results for the second RIR (Vienna) showed worse performance. This discrepancy was attributed to the coefficients and window lengths having been adjusted through informal listening to fit the first RIR (Auditorium) [4]. However, the results showed that the performance of HO-SIRR is better than Ambisonics for first and third order input.

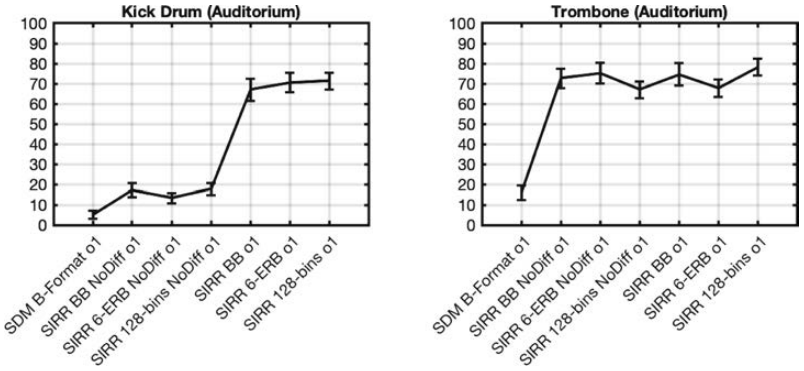

Figure 7: The first test results. [4]

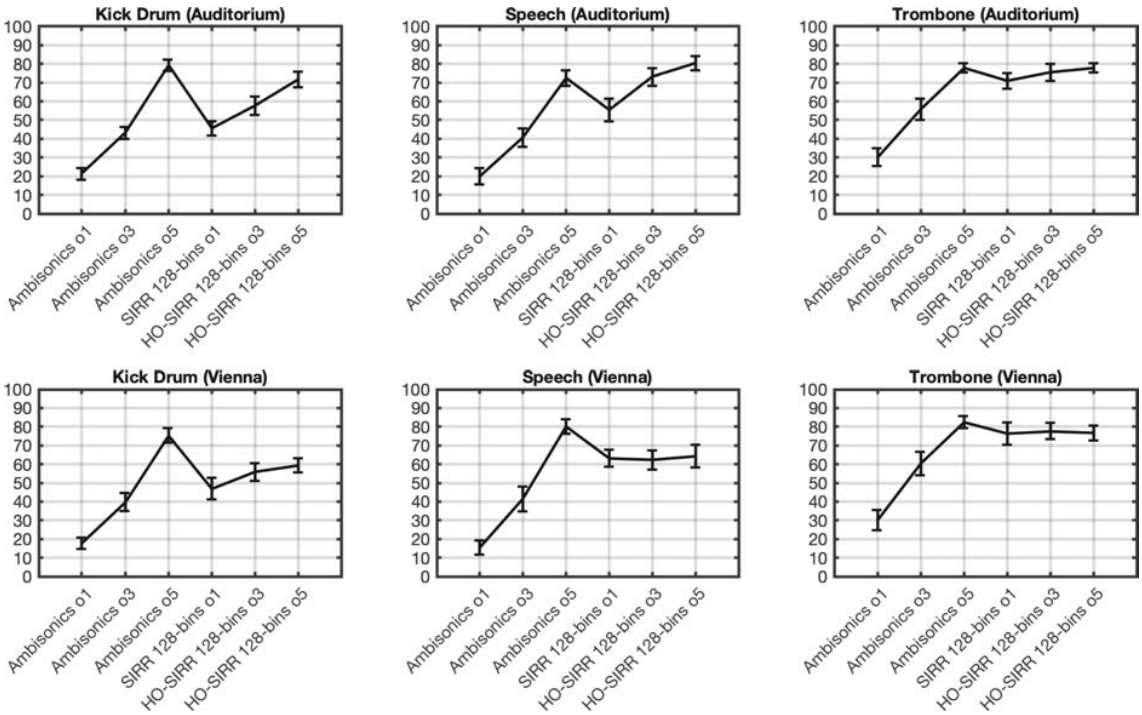

Figure 8: The second test results. [4]

4 Spatial Decomposition Method

Spatial decomposition method was introduced in [14]. The basic idea was that the *direction* of arrival (DOA) could be estimated from the time *difference* of arrival (TDoA) from the signals of a regular microphone array. Additionally, the method was later extended to support similar broad-band PIV DOA detection (SDM B-format), and this was the variant used in [4].

Figure 9 shows the flow diagram for SDM depicted in [4]. Regarding the choise of DOA analysis, in [7], it is mentioned that for open arrays of omnidirectional microphones the TDoA based analysis is preferred, and for SHD input the PIV based DOA detection estimation is preferred. Both of these DOA analysis methods, however, share the same limitation of first order SIRR, of not being able to differentiate the DOA of two temporally closely spaced reflections [7,4]. The analyzed DOA can then be used to either pan the sound to the nearest loudspeaker (relative to DOA) or to multiple loudspeakers using VBAP. The time-dependent equalization is primarily used to mitigate some issues during late reverberation.

Even though originally the method performed better than SIRR in listening tests [14], in [4] it was noted that this was because the authors did not employ an anechoic room for the listening test. Another confounding factor was deemed to be the lower number of

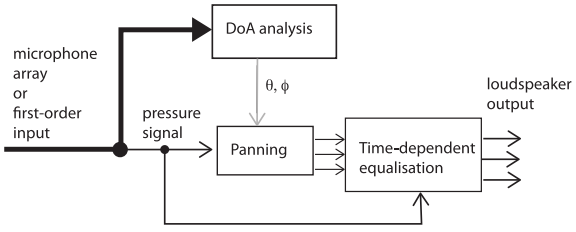

Figure 9: Block diagram of the SDM method in [4]

loudspeakers used in the original listening test. According to [4,7], with higher number of loudspeakers, SDM (using the nearest loudspeaker rendering) produces a sparse sound. This is caused by the limited number of loudspeaker channels used in the reproduction compared to, for example, HO-SIRR where Ambisonics was used for the diffuse stream.

5 Conclusions

Based on the findings of [4], it appears that regarding (HO-)SIRR, there is still room for research on specifying the coefficients and time windows for optimal reproduction quality. When it comes to SDM, in [7] several improvements were said to have been made to the SDM-toolbox implementation to compensate for the issues with roughness and transient signals, that were also found in [4]. However, the last update to SDM-toobox was before the paper was published, in 2018 [15].

The author of this seminar paper declares the fact that they have not had an opportunity to read all the relevant literature on the topic during the time of writing. Therefore, the conclusions might be outdated and flawed.

References

[1] Amir Avni, Jens Ahrens, Matthias Geier, Sascha Spors, Hagen Wierstorf, and Boaz Rafaely. 2013. Spatial perception of sound fields recorded by spherical microphone arrays with varying spatial resolution. *The Journal of the Acoustical Society of America* 133, (2013), 2711–2721. https://doi.org/10.1121/1.4795780

[2] Stéphanie Bertet, Jérôme Daniel, Etienne Parizet, and O. Warusfel. 2013. Investigation on Localisation Accuracy for First and Higher Order Ambisonics Reproduced Sound Sources. *Acta Acustica united with Acustica* 99, (2013), 642–657. Retrieved from https://hal.science/hal-00848764

[3] Sebastian Braun and Matthias Frank. 2011. Localization of 3D Ambisonic Recordings and Ambisonic Virtual Sources. In *International Conference on Spatial Audio (ICSA)*, November 2011. Retrieved from https://www.microsoft.com/en-us/research/publication/localization-of-3d-ambisonic-recordings-and-ambisonic-virtual-sources/

[4] Leo McCormack, Ville Pulkki, Archontis Politis, Oliver Scheuregger, and Marton Marschall. 2020. 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, 5 (June 2020), 338–354. https://doi.org/10.17743/jaes.2020.0026

[5] Juha Merimaa and Ville Pulkki. 2004. Spatial impulse response rendering. In *Proc. of the 7th Intl. Conf. on Digital Audio Effects (DAFX'04), Naples, Italy*, 2004. 29–30.

[6] Juha Merimaa and Ville Pulkki. 2005. Spatial Impulse Response Rendering I: Analysis and Synthesis. *J. Audio Eng. Soc* 53, 12 (2005), 1115–1127. Retrieved from https://www.aes.org/e-lib/browse.cfm?elib=13401

[7] Nils Meyer-Kahlen, Sebastià V. Amengual, and Tapio Lokki. 2022. What the Spatial Decomposition Method can and cannot do. In *ICA 2022 proceedings (Proceedings*

of the ICA congress), 2022. Acoustical Society of Korea (ASK), Korea, Republic of. Retrieved from https://ica2022korea.org/

[8] Ville Pulkki and Juha Merimaa. 2006. Spatial Impulse Response Rendering II: Reproduction of Diffuse Sound and Listening Tests. *J. Audio Eng. Soc* 54, 1/2 (2006), 3–20. Retrieved from https://www.aes.org/e-lib/browse.cfm?elib=13664

[9] Ville Pulkki, Symeon Delikaris-Manias, and Archontis Politis (Eds.). 2017. *Parametric Time-Frequency Domain Spatial Audio*. Wiley-Blackwell, United States. https://doi.org/10.1002/9781119252634

[10] Ville Pulkki. 2007. Spatial Sound Reproduction with Directional Audio Coding. *J. Audio Eng. Soc* 55, 6 (2007), 503–516. Retrieved from https://www.aes.org/e-lib/browse.cfm?elib=14170

[11] Olli Santala, Heikki Vertanen, Jussi Pekonen, Jan Oksanen, and Ville Pulkki. 2009. Effect of Listening Room on Audio Quality in Ambisonics Reproduction. In *Audio Engineering Society Convention 126*, May 2009. Retrieved from https://www.aes.org/e-lib/browse.cfm?elib=14860

[12] Audun Solvang. 2008. Spectral Impairment of Two-Dimensional Higher Order Ambisonics. *J. Audio Eng. Soc* 56, 4 (2008), 267–279. Retrieved from https://www.aes.org/e-lib/browse.cfm?elib=14385

[13] Peter Stitt, Stephanie Bertet, and Maarten van Walstijn. 2014. Off-Centre Localisation Performance of Ambisonics and HOA For Large and Small Loudspeaker Array Radii. *Acta Acustica united with Acustica* 100, 5 (September 2014), 937–944. https://doi.org/10.3813/AAA.918773

[14] Sakari Tervo, Jukka Pätynen, Antti Kuusinen, and Tapio Lokki. 2013. Spatial Decomposition Method for Room Impulse Responses. *J. Audio Eng. Soc* 61, 1/2 (2013), 17–28. Retrieved from https://www.aes.org/e-lib/browse.cfm?elib=16664

[15] Sakari Tervo. 2024. SDM Toolbox. Retrieved from https://www.mathworks.com/matlabcentral/fileexchange/56663-sdm-toolbox

[16] Liu Yang and Xie Bosun. 2014. Subjective evaluation on the timbre of horizontal ambisonics reproduction. In *2014 International Conference on Audio, Language and Image Processing*, 2014. 11–15. https://doi.org/10.1109/ICALIP.2014.7009747