Impulse-response-based 6DOF reproduction of spatial sound

Jackie Lin
Aalto Universtiy
Master's Programme CCIS / AAT

jackie.lin@aalto.fi

Abstract

6 Degrees of freedom (6DOF) reproduction of acoustics aims to provide a dynamic acoustics rendering to a listener that can move in six degrees of freedom. In this review, impulse response-based 6DOF reproduction techniques are reviewed. First, two-point interpolation is covered. This includes a review of mathematical interpolation techniques for two signals, which is the basis of many techniques to interpolate available RIRs to listener perspectives, followed by partial optimal transport for the interpolation between two known RIRs. Following this, multi-point interpolation is reviewed which relies on multiple (three) known listener perspectives. These techniques rely on Ambisonic room impulse responses (ARIRs) which contain spatial direction of incident plane waves that make up the RIR. The extra spatial information about the sound field or sound events (highly related) allows for the use of multiple ARIRs in interpolation. Lastly, neural acoustic fields, a deep learning-based interpolation, is presented which has advantages in dynamic source and listener positions but only applicable to one acoustic scene.

1 Introduction

6DOF spatial sound reproduction delivers immersive, lifelike soundscapes that dynamically adjust to the listener's movements within a given scene. Unlike conventional stereo or surround sound setups, 6DOF spatial audio technology creates an environment where sound objects appear to exist in specific locations relative to the listener, regardless of their position or orientation. This innovative approach ensures that the auditory perception remains consistent and realistic, mirroring the complexities of real-life sound scenarios. Whether walking, jumping, or simply turning around, individuals immersed in a 6DOF spatial sound environment encounter a realistic auditory journey, where every shift in perspective seamlessly aligns with

the spatial positioning of sound sources, heightening the overall sense of presence and immersion. The spatial reproduction of a sound scene that responds to the arbitrary translations and rotations of a listener is referred to as variable listener perspective auralization.

In this seminar paper, room impulse response (RIR)-based 6DOF reproduction methods are reviewed. IR-based 6DOF reproduction auralizes sound sources by convolving dry signals with directional room impulse responses while taking into account variable listener position and orientation. The 6DOF reproduction of ambisonics recordings is not in the scope of this review.

The paper is organized as follows: Section 2 covers traditional mathematics-based interpolations such as linear time-domain interpolation and linear frequency-domain interpolation. It also discusses interpolating between two receiver positions using partial optimal transport which results in an RIR with the desired peak and temporal shifts. Section 3 covers RIR interpolation from multiple RIRs distributed throughout a room. Section 4 covers neural acoustic fields, a deep learning approach for RIR interpolation. Finally, Section 5 summarizes the paper.

2 Two-point Interpolation

2.1 Mathematical Interpolations

This section explains two basic interpolations that interpolate two signals covered in [1]: linear interpolation in the time domain and linear interpolation in the frequency domain. These mathematical interpolations often act as a baseline interpolation for performance evaluation and are used as a component in a greater interpolation scheme.

2.1.1 Linear Interpolation in Time Domain

Linear interpolation of two time-domain RIRs is the linear combination of the two signals weighted by distance criteria between the extrapolated perspective and known perspectives.

Over a 2D rectangular grid of RIRs, bilinear interpolation is performed which consists of three linear interpolations; for a listener perspective at (x_l, y_l) two linear interpolations are performed between $(x_i, y_i)(x_{i+1}, y_i)$, and $(x_i, y_{i+1})(x_{i+1}, y_{i+1})$. These are then linearly interpolated a third time to achieve the desired signal.

Linear interpolation does not work well because it averages sample amplitudes rather than interpolating in the time dimension to make one peak. Figure 1 demonstrates time-domain linear interpolation.

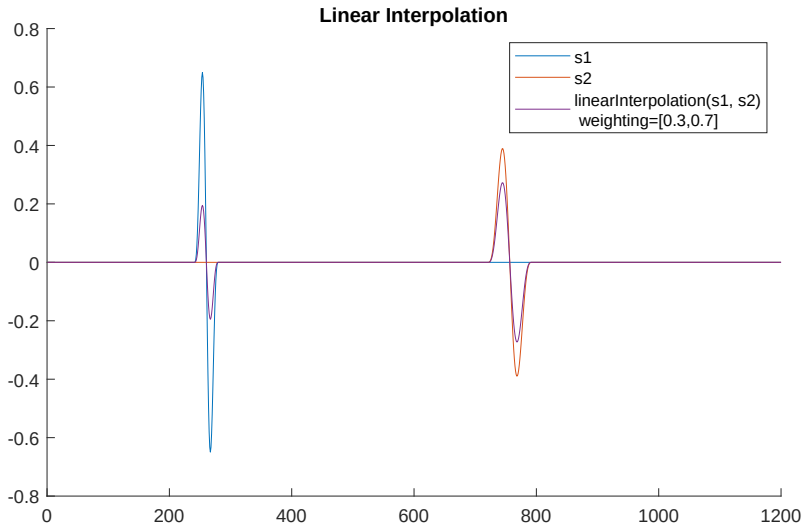

Figure 1: Linear interpolation in the time domain.

2.1.2 Linear Interpolation in Frequency Domain

The RIRs are translated to the frequency domain and interpolated in magnitude and phase. This results in a better interpolation method than linear interpolation in the time domain, as shown in Figure 2. The amplitudes of the two peaks are interpolated and a single peak is shifted in the time domain between the original two peaks.

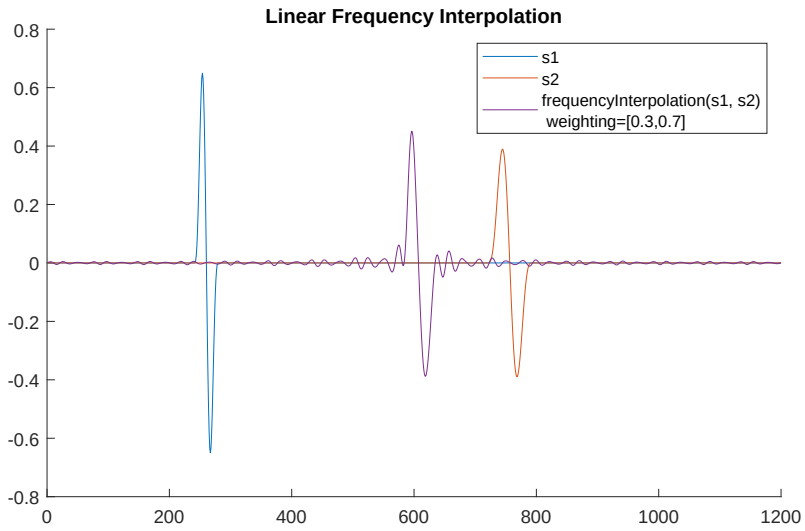

Figure 2: Linear interpolation in the frequency domain.

2.2 Interpolation using Partial Optimal Transport

Partial optimal transport has also been used to interpolate spatial room impulse responses[1] (SRIR) [2], in which a direct correspondence between the direct and early peaks in the available SRIRs is not solved for. Instead, a proposed coupling between reflections is found through partial optimal transport. The method here proposes an interpolation plan for the straight line between two listener locations and a fixed source location, thus only needing two SRIRs. If the problem is extended to more listener positions, there must be some constellation of ground truth SRIRs and multiple partial optimal transport plans.

This method utilizes a geometrical acoustics premise where the direct and early reflections of the SRIR are emitted from virtual sound sources, where each virtual sound source represents a reflection path from source to receiver. Thus the early portion of the SRIR can be represented by a point cloud of virtual sources $\mathcal{P}$. In the implementation presented, the exact locations of the virtual sources in the point cloud are known because the room geometry is known and the SRIRs are simulated. However, to construct the point cloud of a measured SRIR, the SRIR is divided into short segments where each segment represents a virtual source: the virtual source is located in space by estimating the time of arrival (TOA) and direction of arrival (DOA) of that segment.

Then given a fixed sound source location, the interpolation problem between two point clouds $\mathcal{P}$, $\mathcal{Q}$ representing two receiver locations at $\boldsymbol{s}_\mathcal{P}$ and $\boldsymbol{s}_\mathcal{Q}$ is to find point cloud $\mathcal{R}$ at an intermediate receiver location $\boldsymbol{s}_{\mathcal{R},\kappa} = (1-\kappa)\boldsymbol{s}_\mathcal{P} + \kappa\boldsymbol{s}_\mathcal{Q}$ for interpolation parameter κ where $0 \leq \kappa \leq 1$. Figure 3 shows two point clouds for two different receiver locations and the same source location.

The objective in classical optimal transport (OT) theory is to find an optimal coupling matrix $\mathbf{T} \in \mathbb{R}^{N\times M}$ that defines the mass transported between $\mathcal{P}$ and $\mathcal{Q}$, where the number of discrete masses are $N = |\mathcal{P}|$ and $M = |\mathcal{Q}|$. The cost to transport mass – defined here as the effective pressures $\mathbf{p}$ and $\mathbf{q}$ – from one virtual source to another is defined by the squared Euclidian distance between sources

$$C_{i,j} = \|\boldsymbol{x}_{\mathcal{P},i} - \boldsymbol{x}_{\mathcal{Q},j}\|_2^2 \,.$$

The optimal transport plan is found by minimizing the Frobenius inner product of the cost and transport matrices, i.e.,

$$\min_{\mathbf{T}} \langle \mathbf{C}, \mathbf{T} \rangle_F = \min_{\mathbf{T}} \sum_{i=1}^{N} \sum_{j=1}^{M} C_{i,j} T_{i,j} \tag{1}$$

[1]The SRIR is room impulse response where the virtual source locations for each peak in the RIR is known or estimated, for example via extracted virtual source locations from Ambisonic RIRs from the pseudo intensity vector for DOA. It can also be a collection of virtual sources with location information that contribute known energy to the RIR.

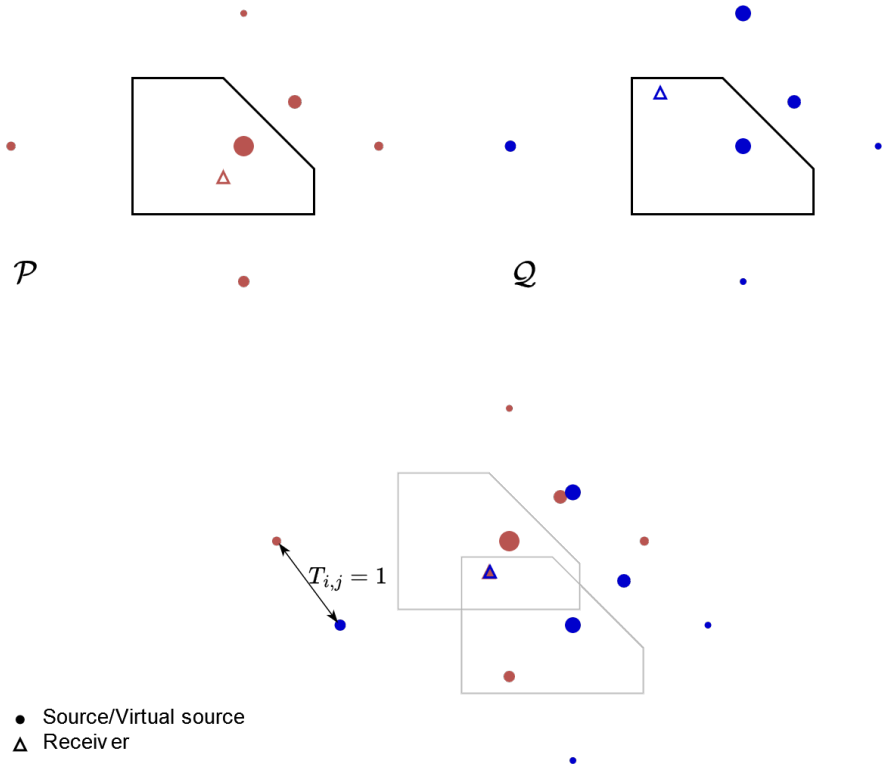

Figure 3: Top row: Two point clouds $\mathcal{P}$, $\mathcal{Q}$ for two source-receiver pairs in the same room. The size of the dot represents the effective pressure of the sound source. Bottom row: $\mathcal{P}$, $\mathcal{Q}$ superimposed with the respective receiver positions at the origin. The partial optimal transport matrix dictates the amount of mass transferred between the points in one cloud to another.

$$\begin{aligned} \text{s.t.} \quad \mathbf{T}\mathbf{1}_M &= \mathbf{p}\,, \\ \mathbf{T}^\intercal \mathbf{1}_N &= \mathbf{q}\,, \end{aligned}$$

to constrain that all the mass of each virtual source p_i in $\mathcal{P}$ is transferred to the virtual sources q_i in $\mathcal{Q}$ and vice versa.

In the case of SRIR point clouds and many other transportation problems, the mass of one point cloud is different from the mass of the other. Hence, the partial optimal transport problem relaxes the mass constraint by utilizing dummy points to the cost matrix and transportation matrix to allow for the introduction or removal of mass in the transportation plan.

The authors utilize a Matlab optimization solver to find the partial optimal transport plan $\bar{\mathbf{T}}$ which fully and continuously provides an interpolation for any receiver location $\boldsymbol{s}_\mathcal{R}$ on the straight-line path between $\boldsymbol{s}_\mathcal{P}$ and $\boldsymbol{s}_\mathcal{Q}$. To construct the interpolated RIR from $\{\bar{\mathbf{T}}, \kappa, \mathcal{P}, \mathcal{Q}\}$, an algorithm to handle three possible mappings (transported mass, vanishing/appearing mass, and moving dummy mass) is proposed. A comparison between interpolated SRIRs and the ground truth SRIRs show that partial OT outperforms linear time-domain mapping and nearest neighbor mapping.

Partial OT for SRIR interpolation is flexible; the method works for non-cuboidal rooms, disappearing and appearing virtual sources (due to occlusion), and provides a continuous interpolation solution for the entire straight path between two receiver locations. However, this interpolation is not a good solution for the late reverberation since the exponential increase of point sources with respect to time. This optimal transport plan however only returns an interpolation between two receiver locations, and future work needs to be done to extend partial OT to a grid of SRIRs.

3 Multiple-Point Interpolation

The second paradigm for 6DOF auralization uses multiple measured or simulated RIRs from multiple perspectives, typically ambisonic room impulse responses (ARIR), to interpolate the RIR at the listener's perspective. This is followed by a review of a multi-perspective ARIR interpolation method [4] and sparse plane wave decomposition method [5].

3.1 Interpolation of Multi-perspective ARIRs

To achieve better interpolation of spatial information from a grid of ARIRs, multiple ARIRs can be used to generate an accurate ARIR at the listener's perspective. Müller and Zotter [4] propose a multi-perspective ARIR interpolation scheme which applies an involved interpolation of a triplet of ARIRs to a new listener perspective.

First, a triplet of ARIRs are selected from a grid of available ARIRs such that the desired listener perspective falls within the triangulation of the selected ARIRs. This selection process is analogous to vector base amplitude panning (VBAP) [6] where a triplet of loudspeakers are selected based on sound source location.

Then, for each selected ARIR, a perspective *extrapolation* is applied to achieve an ARIR at the desired listener perspective: an ARIR $h_i(t)$ with perspective x_i will be transformed to an ARIR $h_d(t)$ at the listener perspective x_d. ARIR extrapolation is done by segmenting the selected ARIR into short time chunks where each segment is modeled as an independent sound event happening at position x_t. Using the known perspective difference between ARIR and listener position, each sound event is extrapolated to the listener perspective by applying the appropriate rotation, gain, and time adjustment, i.e.

$$h_d(t) = G(x_t, x_i, x_d)R(x_t, x_i, x_d)h_i(t + \Delta\bar{t}(x_t, x_i, x_d)) \tag{2}$$

One contribution of this work by Müller et al. [4] is a method of preserving temporal context using a novel quantized time-shift map $\Delta\bar{t}(x_t, x_i, x_d)$ in the extrapolation step.

If the available ARIRs and thus subsequent extrapolated ARIR are first-order, Ambisonic spatial decomposition method (ASDM) by Zaunschirm et al. [8] is applied to enhance the perceived spaciousness by transforming the first-order ARIR to a higher-order Ambisonics signal. ASDM assumes a single time-varying direction as the carrier of the sequence of the broadband sound signal. The direction of arrival (DOA) vector of this carrier is estimated from the smoothed pseudo intensity vector of a band-limited first-order ARIR. The omnidirectional channel of the original ARIR is then encoded to a higher-order Ambisonics signal based on the estimated time-varying DOA vector.

The key contribution of the work is using the constellation of ARIRs to perform joint sound-event localization. The joint sound-event localization is done using arrival times of high-energy ARIR peaks and offers a higher localization accuracy than a single DOA-based localization. First, peak detection and fundamental parameter estimation is performed. Then the method for joint sound-event localization is introduced. It includes a global localization of the direct sound source and triplet based sound event localization of early peaks.

After obtaining the three extrapolated ARIRs, they are linearly *interpolated* (hence "multi-perspective" interpolation) in the time domain to return a final listener perspective ARIR.

3.2 Sound Field Interpolation

Sparse plane wave decomposition generates an ARIR for a listener perspective by representing the sound field as sparse plane waves from a set of measured or simulated ARIRs [5, 7]. In [5] the method presented analyzes the ARIRs in the time-frequency domain and selects time-frequency bins with a strong directional component, representing plane waves in the sound field, for interpolation.

Plane wave decomposition expresses an arbitrary, stationary sound field as a linear combination of plane waves. In the method proposed in [5], the more simplistic sparse plane wave decomposition is used which expresses the sound field as a linear combination of fewer salient plane waves and a diffuse field component.

Similar to [4], three ARIRs from the available ARIRs are selected based on if the listener falls within the triangulation. The short-time Fourier transform (STFT) is applied to the selected ARIRs to obtain a time-frequency representation for each ambisonics channel. Then, a residual energy test identifies which time-frequency bins have a strong direct path component, corresponding to the incidence of a sound wave.

The assignment of time-frequency bins as either plane wave or diffuse occurs. The time-frequency bins with a residual energy metric above a user-set threshold are

selected to represent the plane waves. The unselected time-frequency bins represent the diffuse field.

For the plane wave time-frequency bins, the corresponding amplitude and direction of arrival of plane waves at each ARIR position are obtained. To obtain the amplitude of the plane wave at the listener position, the dominant plane wave components from each ARIR are translated to the interpolation position, which is then weighed based on the distance between ARIR to listener position. To obtain the direction of propagation of the plane wave at the listener position, the directions of the plane waves from each ARIR are also averaged by a linear combination. These result in amplitudes and directions of plane waves at the listener position, which are then used to construct an Ambisonics signal at the listener position.

The diffuse time-frequency bins are linearly combined based on the distance from each ARIR position and listener position. The diffuse component interpolation is added to the plane wave signal to return the interpolated ARIR signal.

An evaluation using real measurements showed that linear interpolation in the spherical harmonics domain provided a more diffuse result than the proposed approach and a less accurate localization of the sound source.

4 Neural Representations

Neural acoustic fields (NAF) [3] is deep neural network that produces RIRs at unseen source-listener pair positions and listener orientations. It is interesting to note that so far, the reviewed 6DOF reproduction techniques allow free listener movement but are only valid for stationary sound sources. With NAF, both the listener and receiver positions can be dynamic. NAF is an acoustics adaption of a computer vision neural network approach, neural representation fields (NeRF) which generates an image of an object from some camera perspective.

NAF utilizes a multilayer perceptron to learn a continuous underlying RIR function of a given scene (one or more acoustically connected spaces) from a recorded or simulated dataset. It generates RIRs from novel, variable source and listener positions. NAF's inputs are $\{\boldsymbol{x}', \boldsymbol{x}, \theta, k, t, f\}$ where source position is $\boldsymbol{x}'$, receiver position is $\boldsymbol{x}$, receiver orientation is θ, left/right channel encoding for binaural RIRs is $k \in \{0, 1\}$, time bin is t, and frequency bin is f. Additionally from source-receiver positions, local geometric conditioning is injected from a feature map $grid(\boldsymbol{x}', \boldsymbol{x})$ learned during training. The model's outputs are the log-magnitude and phase angle of the RIR $\boldsymbol{h}$ at time-frequency bin (t, f). Thus NAF generates an STFT magnitude and phase representation of the RIR.

$$\text{NAF} : (\boldsymbol{x}', \boldsymbol{x}, \theta, k, t, f) \rightarrow (\boldsymbol{h}_{\text{STFT\_mag}}(t, f), \boldsymbol{h}_{\text{STFT\_phase}}(t, f)) \tag{3}$$

The strength of NAF is two-fold, and its benefits come from two areas:

1. NAF is highly general: NAF has a very general architecture that does not depend on any physics-based parameters or tuning. It's training dataset is simply monoaural RIRs or binaural RIRs with source-listener location and orientation metadata.

2. The implicit representation of the acoustic field allows for (a) interpolation that outperforms conventional interpolation techniques and (b) high compression of IRs and fast computation:

   (a) NAF outperforms bilinear and nearest neighbor interpolation. NAF performance was compared to interpolated IRs coded with Advanced Audio Coding (AAC), and Xiph Opus (Opus) formats, and image source method baselines on reverberation time error T_{60} and spectral loss percentage. A listening test with two-alternative forced-choice task showed that 82.38% NAF RIRs were higher quality than AAC-nearest. Figure 4 qualitatively demonstrates that NAF produces a smooth loudness map that approximates the ground truth better than the linear interpolation and nearest neighbor interpolation.

   (b) The implicit representation of the neural field is stored within the parameters of NAF's neural network. Once an NAF model is trained for a scene, generating an IR is a simple as a forward pass through NAF. The NAF model is able to better approximate the ground truth at a fraction of the storage of the impulse responses required to train it.

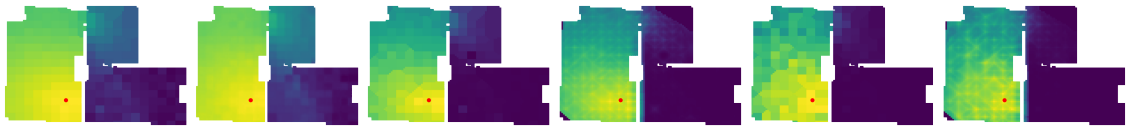

Figure 4: Loudness maps of a multi-room scene. From left to right: Ground-truth nearest neighbor, NAF, AAC nearest, AAC linear interpolation, Opus nearest, Opus linear. Notice how NAF properly captures loudness in the secondary and tertiary room, and coupled room effects at thresholds.

The drawback of NAF is that each model only represents that specific acoustic space, so a trained NAF is not able to generalize to novel environments. The other multi-perspective ARIR methods described above are based on a strong model of sound fields, and thus are applicable to any set of ARIRs.

In the implementation of NAFs, it took on average 6 hours to train a model on one scene. Additionally, the training dataset was 90% of the available RIRs for that room, which is very high. It is unclear if linear interpolation or nearest neighbor interpolation is perceptually the same as the NAF interpolation, since the number of available RIRs is so large.

5 Summary

To summarize, mathematical interpolations such as linear interpolation are useful as a baseline for comparison or when integrated into a larger interpolation framework. At its core, the mathematical interpolation is the linear combination of two signals. Partial optimal transport for spatial RIR interpolation finds the virtual source contributions for any listener position on the straight line between two listener positions of known SRIRs. This is an interesting framing of RIR interpolation but its application is limited and more work must be done to interpolate the SRIRs in a plane.

More practical interpolations such as the multi-perspective ARIR interpolation and the sparse plane wave decomposition ARIR interpolation have been proposed which utilize spatial information in these ARIRs and some underlying model of sound propagation. For a set of ARIRs with fixed source location and varying listener locations, sound events can be localized in space using the spatial qualities in the ARIR and then the resulting ARIR at the listener location can be found. In the multi-perspective interpolation, the existing ARIRs are segmented in time where each segment corresponds to a sound event occurring at some distance from the recorded perspective. The spatial information of this sound event can be used to calculate the sound event as it is heard at the listener perspective. In the sparse plane wave decomposition interpolation, the ARIR is segmented into time-frequency bins which correspond to plane waves arriving from some direction. The plane waves here originate from the sound events in the multi-perspective interpolation: these are the same thing.

For all the methods described, the source location is static while the listener moves. However, a recent deep learning approach, neural acoustic fields, has allowed for the smooth interpolation of RIRs that accommodate dynamic source and listener movement. This is a very powerful and exciting direction, however, an NAF is trained only on one acoustic scene so a new model must be retrained from scratch for every new scene. This paper reviews some exciting and interesting methods for 6DOF reproduction.

References

[1] García Gómez, V. Técnicas de interpolación en sistemas de sonido espacial avanzados.

[2] Geldert, A., Meyer-Kahlen, N., and Schlecht, S. J. Interpolation of spatial room impulse responses using partial optimal transport. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2023), IEEE, pp. 1–5.

[3] Luo, A., Du, Y., Tarr, M., Tenenbaum, J., Torralba, A., and Gan, C. Learning neural acoustic fields. *Advances in Neural Information Processing Systems 35* (2022), 3165–3177.

[4] Müller, K., and Zotter, F. Auralization based on multi-perspective ambisonic room impulse responses. *Acta Acustica 4*, 6 (2020), 25.

[5] Olgun, O., Erdem, E., and Hacihabiboğlu, H. Sound field interpolation via sparse plane wave decomposition for 6dof immersive audio. In *2023 Immersive and 3D Audio: from Architecture to Automotive (I3DA)* (2023), IEEE, pp. 1–10.

[6] Pulkki, V. Virtual sound source positioning using vector base amplitude panning. *Journal of the audio engineering society 45*, 6 (1997), 456–466.

[7] Wang, Y., and Chen, K. Translations of spherical harmonics expansion coefficients for a sound field using plane wave expansions. *The Journal of the Acoustical Society of America 143*, 6 (2018), 3474–3478.

[8] Zaunschirm, M., Frank, M., and Zotter, F. Brir synthesis using first-order microphone arrays. In *Audio Engineering Society Convention 144* (2018), Audio Engineering Society.