Auralization of Measured Room Transitions in Virtual Reality

Citation for published version:
McKenzie, T, Meyer-Kahlen, N, Hold, C, Schlecht, SJ & Pulkki, V 2023, 'Auralization of Measured Room Transitions in Virtual Reality', *Journal of the Audio Engineering Society*, vol. 71, no. 6, pp. 326-337. https://doi.org/10.17743/jaes.2022.0084

Digital Object Identifier (DOI):
10.17743/jaes.2022.0084

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Journal of the Audio Engineering Society

Auralisation of measured room transitions in virtual reality

THOMAS MCKENZIE,[*,1, 2] **NILS MEYER-KAHLEN,**[2] *AES Student Member* **CHRISTOPH HOLD,**[2] *AES Student Member* **SEBASTIAN J. SCHLECHT**[2, 3] *AES Member* **AND VILLE PULKKI**[2] *AES Fellow*

[1] *Acoustics and Audio Group, Reid School of Music, University of Edinburgh, United Kingdom*
[2] *Acoustics Lab, Department of Information and Communications Engineering, Aalto University, Espoo, Finland*
[3] *Media Lab, Department of Art and Media, Aalto University, Espoo, Finland*

To auralise a room's acoustics in six degrees-of-freedom (6DoF) virtual reality (VR), a dense set of spatial room impulse response (SRIR) measurements is required, so interpolating between a sparse set is desirable. This paper studies the auralisation of room transitions by proposing a baseline interpolation method for higher-order Ambisonic SRIRs and evaluating it in VR. The presented method is simple yet applicable to coupled rooms and room transitions. It is based on linear interpolation with RMS compensation, though direct sound, early reflections and late reverberation are processed separately, whereby the input direct sounds are first steered to the relative direction-of-arrival before summation and interpolated early reflections are directionally equalised. The proposed method is first evaluated numerically, which demonstrates its improvements over a basic linear interpolation. A listening test is then conducted in 6DoF VR, to assess the density of SRIR measurements needed in order to plausibly auralise a room transition using the presented interpolation method. The results suggest that, given the tested scenario, a 50 cm to 1 m inter-measurement distance can be perceptually sufficient.

0 Introduction

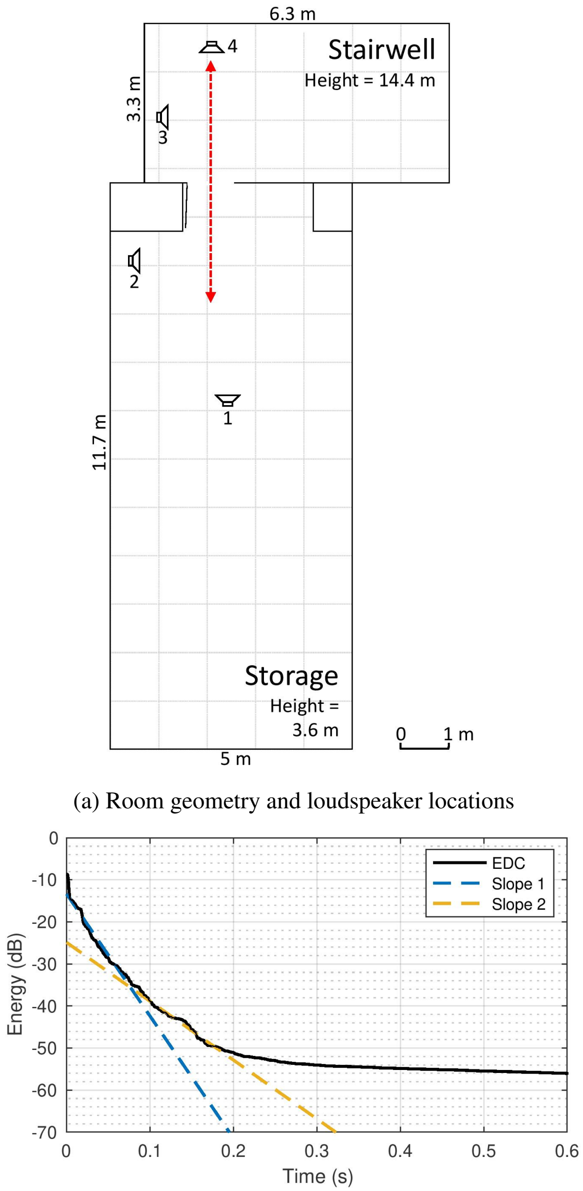
One way to realise virtual acoustics rendering for six degrees-of-freedom (6DoF) immersive experiences is to use room acoustic measurements. The room impulse response (RIR) captures the reverberation of a space. RIRs measured with spherical microphone arrays, which may use the principles of Ambisonics to encode microphone signals into spherical harmonics (SHs) [1], are referred to as spatial room impulse responses (SRIRs). These allow for greater flexibility post measurement, as they can be analysed and modified directionally, and can be reproduced over both loudspeaker arrays and headphones.

Recent literature has investigated how RIRs change with different receiver positions inside a single room for VR applications [2, 3]. While it is possible to interpolate between measured RIRs [4, 5], the perceptual requirements for inter-measurement distance (IMD) have been found to vary with auditory stimuli, whereby sounds with limited frequency bandwidth can forgive larger distances between measurements [6], and the greater diffuseness of late reverberation allows for different IMDs for distinct parts of the impulse response [2]. However, given the greater complexity of coupled room acoustics [7, 8, 9], the perceptual requirements for IMD may differ for room transitions than from those of single room acoustics. Fig. 1 presents a typical room transition scenario.

In a previous study on the acoustics of room transitions [10], some clear trends were observed: direct-to-reverberant ratio decreases when the source and receiver are in opposing rooms with no continuous line-of-sight (CLOS) between the source and receiver (such as loudspeakers 2 and 3 in Fig. 1), and increases for less reverberant rooms. These effects are greater when the difference in reverberation between the two rooms is larger, and change depending on the source position. Directional analysis showed that the reflection patterns are generally consistent in each room, but become more intricate in the region around the coupling aperture. Additionally, it showed the presence of strong reflections, sometimes with a greater amplitude than the occluded direct path, especially around the coupling aperture.

For 6DoF rendering of a sound scene with Ambisonic SRIRs at multiple listener positions in space, a convolution implementation that can switch between measurements in real-time is needed, followed by an auralisation method such as binaural rendering. Comparisons of auralisations with real loudspeakers require particular care in the equali-

[*] To whom correspondence should be addressed, e-mail: thomas.mckenzie@ed.ac.uk



(a) Room geometry and loudspeaker locations

(b) EDC (energy decay curve) for loudspeaker 2 at 150 cm inside storage space

Fig. 1: Room geometry and loudspeaker locations of the room transition, and an energy decay curve illustrating the double-slope decay of the room transition, of the Storage to Stairwell transition in [10]. Measurements denoted by dashed arrow; loudspeaker numbers 1 and 4 retain a continuous line-of-sight between the loudspeakers and microphone for all measurement positions, 2 and 3 feature occlusion at some measurement positions.

sation of the auralisation [11, 12, 13]. An alternative workflow is to interpolate between running signals of multiple Ambisonic receivers in real-time [14, 15, 13], though this study will focus on the former workflow based on SRIR convolution.

Interpolation of single-channel RIRs has been approached in different ways in the past: 1) Dynamic time warping [16], where the time axes of the nearest RIRs are stretched until they align; 2) Modal interpolation using a general solution to the Helmholtz equation [17], which is effective for non-uniform spatial distributions of RIRs at low frequencies; and 3) A combination of plane wave decomposition and time-domain equivalent source methods [4].

For interpolating SRIRs, sometimes called directional RIRs (DRIRs), a first-order interpolation method is presented in [18], which separates input SRIRs into specular parts, which are the direct sound and early reflections, and the diffuse parts. These are interpolated separately, where the specular parts are interpolated individually using direction-of-arrival (DoA) estimations. In [5], a similar method is presented for early reflection interpolation between the nearest three receivers, with simpler interpolation of residual signals. In [19], a Gaussian regression model method is presented, which offers not only interpolation but also extrapolation. However, these methods are all intended for use in a single room and may thus not cope with the more complex acoustics of coupled rooms.

Interpolation between coupled room SRIRs is likely to be a more demanding task than for SRIRs inside the same room, as many acoustical properties may vary significantly between measurements, including the energy decay [20] and direct-to-reverberant ratio [10]. Occluded line-of-sights between the source and receiver can lead to large differences between estimated source positions; methods that attempt to determine the sound source and early reflection positions through triangulation techniques would therefore be inappropriate.

This paper presents a perceptually informed interpolation method for higher-order Ambisonic SRIRs, which utilises interpolation of the direct sound steered to the estimated DoA and linear interpolation with directional equalisation in the early part of the response, and relative RMS matching late reverberation interpolation. This method is first evaluated numerically, in comparison to a basic linear interpolation method, using RMS and DoA error metrics. A listening test is then conducted in VR to assess the perceived quality of interpolating between coupled room SRIRs with varying IMD. The results of the test are used to suggest an appropriate IMD for accurate reproduction of room transitions when using a relatively simple interpolation method.

This paper is laid out as follows: Section 1 details the proposed method for SRIR interpolation, including the separate methods used for direct sound, early reflections and late reverberation. Section 2 then evaluates the proposed method, first numerically and then perceptually. Finally, Section 3 presents concluding remarks and proposes further work. Links are provided to the room transition SRIR dataset, a MATLAB implementation of the presented the interpolation method, and an open-source virtual studio technology (VST) plugin for 6DoF convolution.

1 PERCEPTUALLY INFORMED SRIR INTERPOLATION

This section describes the method of perceptually informed interpolation between SRIRs. The method is designed for 3D, 2D, or 1D (along a line) sets of SRIRs, and is therefore appropriate for 6DoF rendering workflows. The maximum SH order is denoted in this paper as N, with the order of an individual SH component denoted by n and the degree denoted by m. In this paper, the Ambisonic Channel Numbering (ACN) and semi-normalised (SN3D) conventions are employed.

Consider a set of J points at Cartesian coordinates $\mathcal{P}_J \subset \mathbb{R}^3$. At each of these points a directional room impulse response is measured, which was encoded to the SH domain and is denoted as $\mathbf{h}_j(t) \in \mathbb{R}^{(N+1)^2}$, where $1 \leq t \leq T$ denotes time in samples and the number of measurements is $1 \leq j \leq J$. These responses are then interpolated to a dense set of $I > J$ points at positions $\hat{\mathcal{P}}_I \subset \mathbb{R}^3$. The distance between a point from the set of measurement points $\mathbf{p}_j \in \mathcal{P}_J$ and a point from the set of interpolation points $\hat{\mathbf{p}}_i \in \hat{\mathcal{P}}_I$ is denoted as

$$v_{i,j} = \|\hat{\mathbf{p}}_i - \mathbf{p}_j\|_2, \tag{1}$$

where $\|\hat{\mathbf{p}}_i - \mathbf{p}_j\|_2$ denotes the Euclidean distance. With the definition of the distance, it is possible to find the subset of $J' = 2^D$ measurement points, which contains the measurements closest to any interpolation point $\hat{\mathbf{p}}_i$, where D is the dimensionality in which the measurement points are arranged. Therefore, a 1D set of SRIRs on a line will have two nearest measurements; a 2D dataset of SRIRs in a grid will have four nearest measurements, and a 3D dataset will have eight nearest measurements. This gives a subset of nearest points $\mathcal{P}_{J'}^{(i)} \subset \mathcal{P}_J$ for each interpolation point i. As all steps described next are carried out for each interpolation point, the index i is omitted for readability.

1.1 Direct Sound

Input SRIRs are time-aligned and truncated to begin at the onset. In this study, the direct sound is taken as the first 4.17 ms of the input SRIRs (200 samples at 48 kHz), though this value is adjustable. The DoA of the direct sound in each input SRIR is first estimated using the time-averaged pseudointensity vector, $\mathbf{i} \in \mathbb{R}^3$, which is derived from the first-order components as

$$\mathbf{i} = \frac{1}{200}\sum_{t=1}^{200}[h_w(t)h_x(t),\ h_w(t)h_y(t),\ h_w(t)h_z(t)]^{\mathrm{T}}, \tag{2}$$

where superscript $^{\mathrm{T}}$ denotes transposition. Note here that h_w is the omnidirectional SH channel of the SRIR, and h_x, h_y, h_z are the respective x, y, and z axis dipoles. This should also not be confused with h_j, which refers to an entire SRIR at measurement position j.

Normalising the intensity estimate

$$\hat{\boldsymbol{\theta}} = \frac{i}{\|i\|}, \tag{3}$$

provides a direct sound direction $\hat{\boldsymbol{\theta}} \in \mathcal{S}^2$ for each measurement position i. For a non-occluded source, a geometrically correct way would be to estimate the sound source location based on the direct sound DoAs observed at the measurement positions. This could be done by finding the point that is closest to all lines along the DoAs, ideally their intersection point. Then, the direct sound direction could be computed at the interpolated point. However, in coupled rooms, where the sound source may potentially be occluded at one or more measurement positions (see for example loudspeakers 2 and 3 in Fig. 1a), this procedure would be inappropriate. Between two measurements, the location of the first sound energy will change and such geometrical solutions could give arbitrary results. Therefore, an approximation algorithm is proposed in this study to estimate the sound source location. The direct sound direction at each interpolation point is set to

$$\hat{\boldsymbol{\theta}} = \sum_{j'=1}^{J'} \theta_{j'} g_{j'}, \tag{4}$$

where $g_{j'}$ are distance weights obtained from the inverse distances between the interpolated positions and the nearest measurement positions, i.e.,

$$\tilde{g}_{j'} = \frac{1}{v_{j'}}, \tag{5}$$

which are normalised

$$g_{j'} = \frac{\tilde{g}_{j'}}{\sum_{j'=1}^{J'} \tilde{g}_{j'}}. \tag{6}$$

For interpolation, the direct sounds of the nearest input SRIRs are first individually rotated to the calculated target direction $\hat{\boldsymbol{\theta}}$. They are then converted into the frequency domain, magnitude weighted based on the gains $g_{j'}$, and the spectra are then summed. To illustrate this, Fig. 2 presents the original direction of the direct sound of two input SRIRs (from a 1 m IMD subset of LS 2 in the Storage to Stairwell transition [10]) in blue (the corresponding plots for the whole dataset are shown in Fig. 3). They are then rotated to the calculated target direction $\hat{\boldsymbol{\theta}}$, as shown in red. Finally, the interpolated SRIR is shown to follow the same direction, as shown in yellow.

To more generally illustrate the DoA rotation of the direct sound interpolation, Fig. 3 presents the DoA of the direct sound of each measurement in the dataset of loudspeaker 2, first for the original 5 cm inter-measurement distance (IMD), then for a 100 cm IMD subset, then for interpolations of the subset using both the linear and proposed methods. The linear interpolation produces more abrupt changes in direction in contrast to the smooth evolution achieved by the proposed method.

The RMS of each frequency band of the interpolated SRIR direct sound is equalised to match the sum of the RMS of each frequency band of the individually gain-weighted nearest SRIRs. This equalisation is performed in equivalent rectangular bandwidth (ERB) frequency bands [21]. In this study, 48 frequency bands are employed with the lowest frequency at 10 Hz, which approximates to 1/3rd octave bands. The interpolated direct sound is then transformed into the time-domain and amplitude normalised based on the gain weighted RMS of the nearest measurements, whereby the RMS of an SRIR is calculated in this paper as

$$\mathrm{RMS}_{h_j} = \sqrt{\frac{1}{T}\sum_{1}^{T}[h_j(t)^2]}, \tag{7}$$

where for the direct sound, $T = 200$ samples. The RMS correction value g_{RMS} is then calculated as

$$g_{\mathrm{RMS}} = \sum_{1}^{J'} \mathrm{RMS}_{h_{j'}} g_{j'}. \tag{8}$$

This procedure ensures that the effect of the sound source directivity is accounted for at the interpolated position.

Fig. 4a compares the broadband RMS level (omnidirectional channel) of the reference and two interpolated SRIR datasets' direct sound (the first 200 samples). The interpolated sets are generated by interpolating from a subset of the reference dataset with an IMD of 100 cm, for loudspeaker 2 (see again Fig. 1a), where a linear interpolation is simply defined as a distance-weighted summation of the two nearest measurements. Both methods perform comparably.

1.2 Early Reflections

For the early reflection interpolation, firstly the transition time t_{EL}, which is the cutoff between early reflections and late reverberation (sometimes referred to as mixing time [22]) is calculated separately for each input SRIR based on the energy decay curve passing a set threshold value [23]. The omnidirectional channel of each SRIR is first bandpass filtered at 1 kHz, then normalised to a maximum amplitude of 1, and Schroeder integration is used to obtain the energy decay curve (EDC) [24]:

$$D(t) = \int_t^{\infty} h^2(\tau) d\tau. \tag{9}$$

In this study, values of t_{EL} are calculated as the time t when $D(t) = D(1)/10$, rounded to the nearest 1000 samples, which generally fall between 80ms and 250ms for the room transition dataset [10]. This is within typical early reflection cutoff times reported in the literature [25, 23, 22, 26], albeit on the higher end. Values of t_{EL} are calculated separately for each set of nearest measurements, such that if some input measurements are in a dry room and some are in a reverberant room, the transition times will accurately reflect this. For each interpolated SRIR, the early reflections of the nearest input SRIRs are first windowed from 200 samples (the end of the direct sound) to the transition time t_{EL}.

The early reflections are interpolated and equalised at different directions on the sphere by using a process of beamforming and reconstruction, as detailed further in [27, 28]. For this, the measured SH domain responses are analysed with a set of max-$\mathbf{r}_{\mathrm{E}}$ beams directed to a dense set of L uniformly arranged directions on a t-design with steering angles $\mathbf{\Theta}_{\mathrm{t}}$ [29],

$$\bar{\mathbf{h}}_j(t) = \frac{4\pi}{L}\mathbf{Y}_N^{\top}\mathrm{diag}_N\{w_n\}\mathbf{h}_j(t), \tag{10}$$

where $\mathbf{Y}_N \in \mathbb{R}^{(N+1)^2 \times L}$ is a matrix of real spherical harmonics evaluated at directions $\mathbf{\Theta}_{\mathrm{t}}$, and $\mathrm{diag}_N\{w_n\}$ is a diagonal matrix of beamforming weights establishing the desired beam-pattern, with one unique weight for all SH components belonging to each order. The t-design with the least number of points that fulfills $t \geq 2N+1$ is selected. For the fourth-order responses used in this test for example, the t-design has 48 points. The beamformer output signals are weighted with the distance weights and summed together

$$\bar{\mathbf{h}}(t) = \sum_{j'} g_{j'}\bar{\mathbf{h}}_{j'}(t). \tag{11}$$

Next, the summed signals are equalised to match the weighted sum of the magnitude spectra in each direction. Equalisation is needed to rectify any comb-filtering artefacts that may arise from the summation of correlated signals. Comb-filtering is most apparent when the SRIRs to be interpolated are a greater distance apart. Each L beamformed signal is equalised separately, such that colouration is removed in each direction. The alternative would be to use one direction-independent equalisation filter, in which case the beamforming and reconstruction operation would become obsolete. For each ERB band, the target RMS is a sum of the RMS of each amplitude-weighted nearest SRIR beam divided by the current RMS of the interpolated beam. An equalisation curve is then calculated by linear interpolation of each ERB band target RMS, between 20 Hz and 20 kHz. After the directionally equalised responses for every directional response in $\bar{\mathbf{h}}^{(\mathrm{EQ})}$ are obtained, they are brought back in the SH domain [28] using

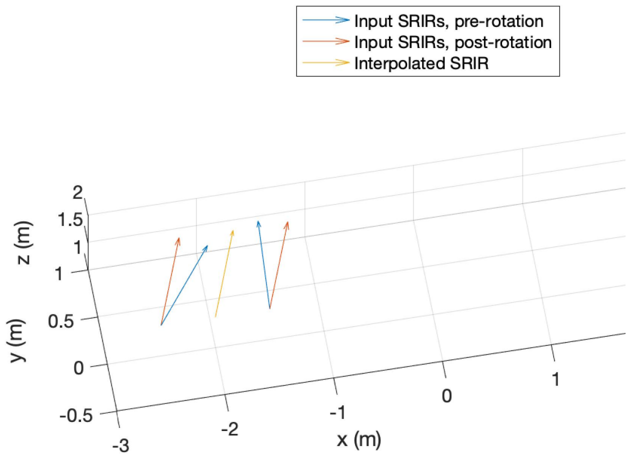
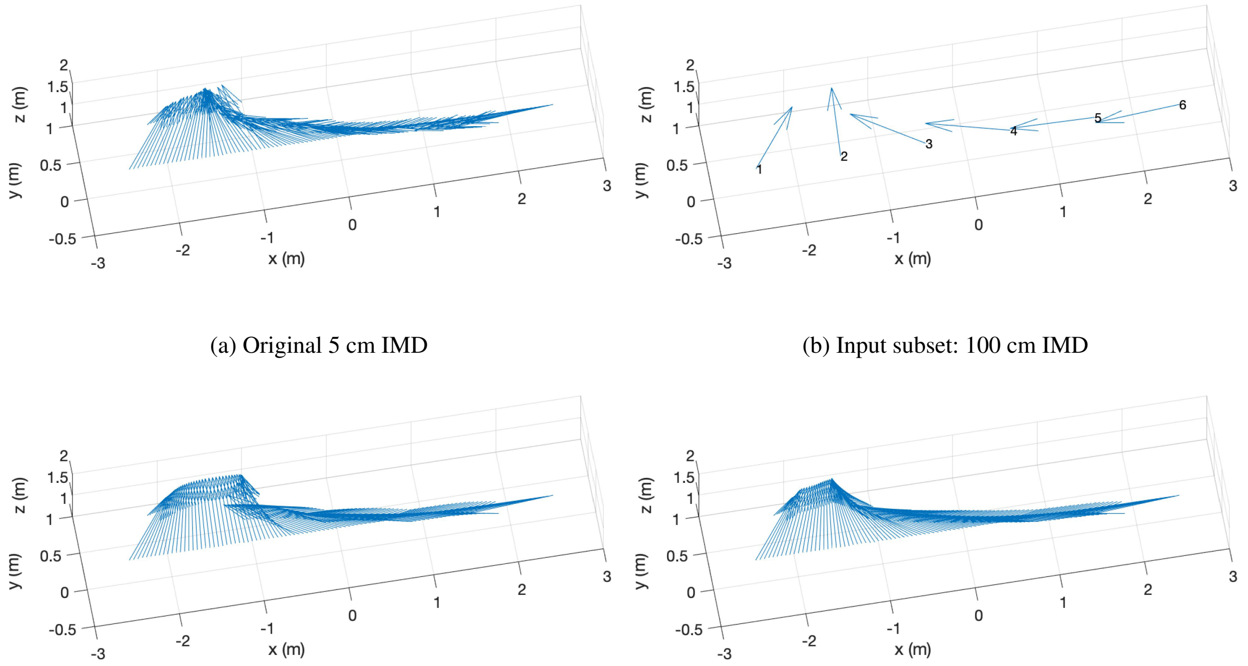


Fig. 2: Illustration of the directional interpolation. Two input SRIRs at -2.5 m and -1.5 m, for an interpolation position at -2.0 m. Direction-of-arrival of the direct sound of two input SRIRs, both pre- and post-rotation to the interpolated direction $\hat{\theta}$, with the direction of the interpolated SRIR.



(a) Original 5 cm IMD

(b) Input subset: 100 cm IMD

(c) Interpolated from 100 cm IMD - linear

(d) Interpolated from 100 cm IMD - proposed

Fig. 3: Direction-of-arrival of the direct sound of each measurement for LS 2, where IMD denotes inter-measurement distance.

$$\mathbf{h}(t) = \mathrm{diag}_N \left\{ \frac{1}{w_n} \right\} \mathbf{Y}_N \bar{\mathbf{h}}^{(\mathrm{EQ})}(t). \tag{12}$$

Fig. 4b compares the overall RMS level of the reference and two interpolated datasets early reflections (from 200 samples to 0.1 s) calculated using (7). The interpolated datasets are again made from a subset of the reference dataset with an IMD of 100 cm, for loudspeaker 2 (see again Fig. 1a). The level drop off between input measurements in the linear interpolation method is evident here, causing a dip in RMS as the distance from a measurement increases, whereas this is mitigated in the proposed method.

1.3 Late Reverberation

The late reverberation interpolation follows much of the same method as used for the early reflections, but without the beamforming, which was deemed perceptually unnecessary. Fig. 4c compares the RMS level of the reference and two interpolated datasets late reverberation (from 0.1 s onwards) calculated using (7). The interpolated datasets are as in Sections 1.1 and 1.2. For the linear interpolation, the level drop off is present here as in the early reflections, while the proposed method follows a smooth trajectory between the RMS of each measurement.

1.4 Construction of the Interpolated SRIRs

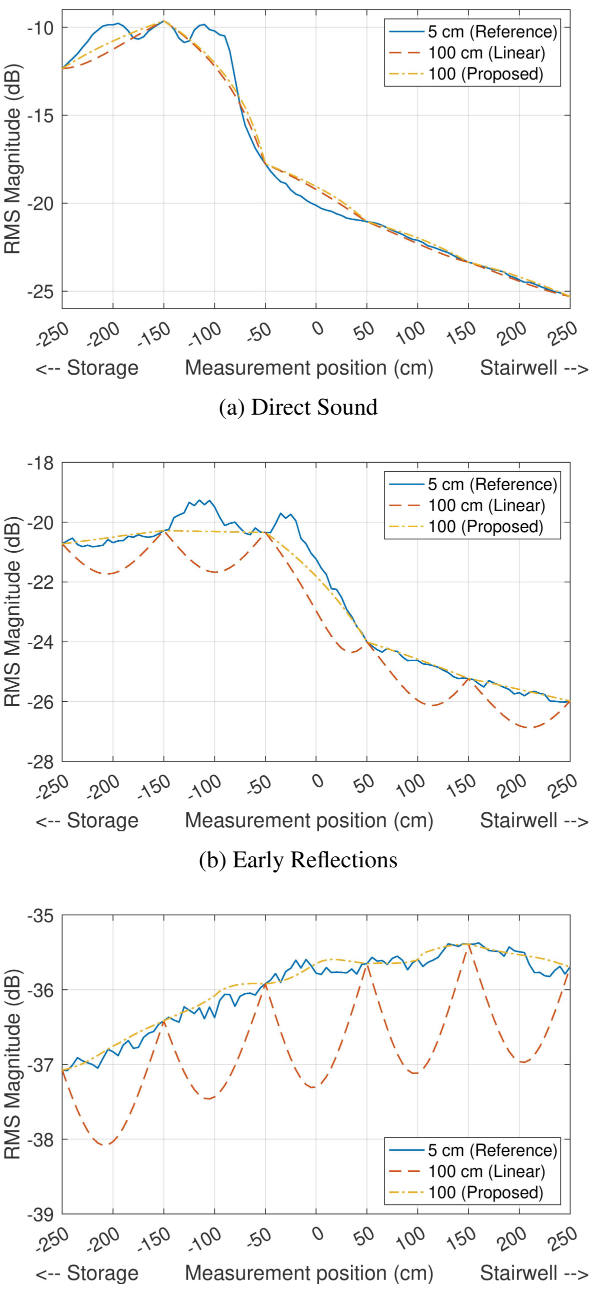
The final interpolated SRIRs are a sum of the interpolated direct sound, early reflections and late reverberation, with cosine-shaped amplitude windows used to fade between sections: 20 samples for direct sound to early reflections, and 10 ms for early reflections to late reverberation (both values are configurable). The interpolated set of SRIRs can then be saved as a spatially oriented format for acoustics (SOFA) file [30], in the same format as the input set, which makes it directly compatible with the Sparta 6DoFconv convolution plugin for auralisation.

2 EVALUATION

This section details the evaluation of the interpolation algorithm, which was carried out both numerically and perceptually. The set of measurements used in the evaluation was the *storage to stairwell* measurements from the room transition dataset of SRIRs at $N = 4$ [10][1].

Fig. 1a presents the room geometry and loudspeaker positions of the measurements, with four loudspeakers: two in each room; for which one retains a continuous line-of-sight (CLOS) between the source and receiver for all receiver positions, and one without CLOS. Fig. 1b shows the EDC, calculated using (9), for loudspeaker 2 at receiver position 100 cm, which is 150 cm inside the storage space. The EDC illustrates the double-slope nature of the energy decay, caused by the combination of the reverberation times of the two rooms, whereby the amplitude of each room's single-slope decay is the only thing that is considered to change with receiver position [31, 20].

[1] `http://doi.org/10.5281/zenodo.4095493`



(a) Direct Sound

(b) Early Reflections

(c) Late Reverberation

Fig. 4: Broadband root-mean-square (RMS) magnitude of the SRIRs (omnidirectional channel) for LS 2 comparing the reference measurements to both linear and the proposed interpolation methods using an IMD of 100 cm.

To assess the interpolation, test datasets of SRIRs were calculated from the original dataset of measured SRIRs, which has a 5 cm inter-measurement distance. This was done by first creating subsets of the original dataset with inter-measurement distances (IMDs) of 10 cm, 25 cm, 50 cm, 100 cm, 250 cm and 500 cm (where the 500 cm case uses just two SRIRs - one at either end), and then interpolating to the same 5 cm resolution as the test dataset. This was repeated for the four loudspeaker positions, for both the proposed and linear interpolation methods.

2.1 Numerical Evaluation

2.1.1 Direction-of-arrival Error

Next, the DoA of the room transition SRIRs was estimated first for the original dataset (with an IMD of 5 cm), and then for the test sets of SRIRs calculated from interpolation of the original dataset with greater IMDs. DoA was estimated above 3 kHz, due to the order-dependent filtering necessary for higher order spherical microphone arrays [32], using a steered response power method, directing a fourth-order hyper-cardioid beamformer (also known as normalised plane wave decomposition), that calculates the power at each chosen location on the sphere [33]. The power was calculated at one degree resolution, which reveals the direct sound and loudest early reflections.

To illustrate the DoA of the interpolated SRIR sets, the horizontal DoA for all measurement positions of LS 2 (in the storage space, no CLOS between the source and receiver) was estimated in five degree resolution for seven arrivals, referring to the direct sound and loudest early reflections. This is presented in Fig. 5, first for the original SRIRs and then for the interpolated SRIRs from 100 cm IMD, both with the linear and proposed interpolation methods. Azimuth is denoted in degrees where a positive increase moves anticlockwise, and colour intensity is normalised separately for each measurement to the maximum power detected in that measurement, in order to illustrate the relative intensity of the dominant source direction to the other reflections. The overall trends are generally reproduced with both interpolation methods, though the direct sound is more accurate with the proposed method. This is especially evident between -150 and -50 cm.

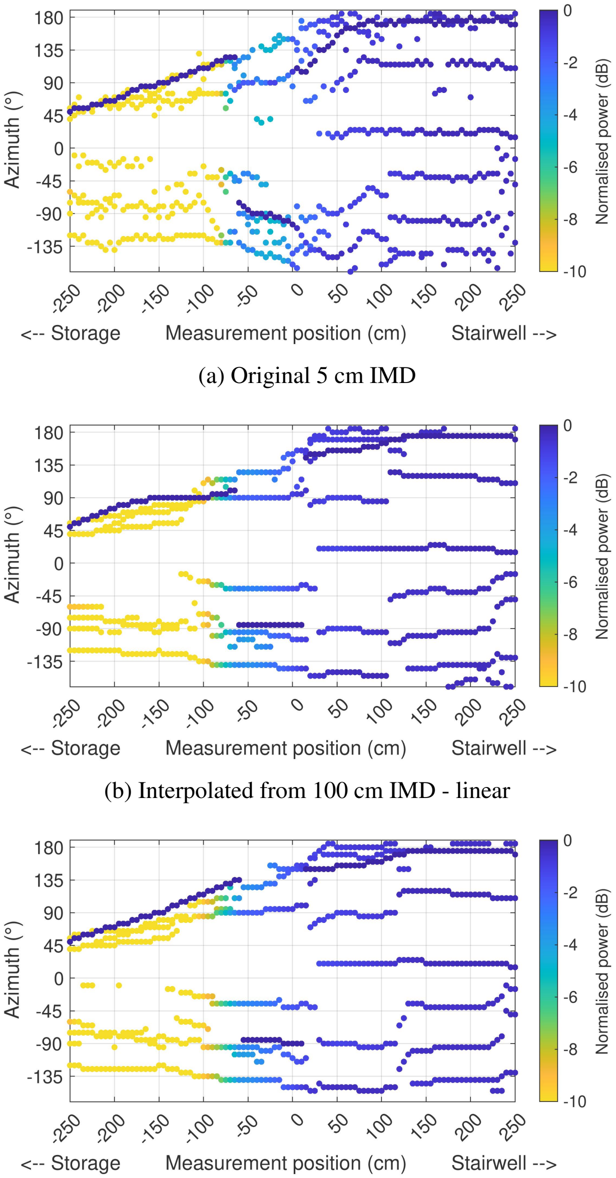
Though the proposed method produced more accurate interpolation of the direct sound, both methods performed similarly for the early reflections. To investigate the DoA error for more IMD values, the DoA was then estimated for just the direct sound (first 200 samples), for the reference dataset and the interpolated datasets at all IMDs. The error in direct sound DoA was calculated as the difference in azimuth angle between the DoAs calculated from the reference dataset and the interpolated datasets, for all 101 measurements. A single azimuth error value E_θ for each interpolated dataset was then calculated as the mean of the 101 absolute differences in estimated azimuth. Table 1 presents the results. In general, E_θ increased with higher IMD, which is expected, and the proposed method produced lower direct sound DoA error than the linear method.

Some interesting results emerge when considering the differences between LS 1 and 2, in the storage space, and LS 3 and 4, in the stairwell. LS 3 and LS 4 have considerably lower E_θ for the low IMD sets, which may be explained by the higher reverberation time of the stairwell, leading to higher energy throughout the room transition. The E_θ significantly jumps at IMD = 500, suggesting the interpolation method is unable to accurately reconstruct the room transition acoustics at this distance, whereas E_θ is significantly higher at IMD = 250 for the linear method.

2.1.2 RMS Error

To investigate the effect on RMS with other loudspeaker positions and IMDs, Table 2 shows the mean RMS error E_{RMS} in dB between the 101 SRIRs of the reference datasets and the test datasets over the entire SRIR response, calculated as

$$E_{\mathrm{RMS}} = \frac{1}{J}\sum_{1}^{J}\left|10\log_{10}\left(\frac{\mathrm{RMS}_{h_j}^{\mathrm{Ref}}}{\mathrm{RMS}_{h_j}^{\mathrm{Int}}}\right)\right|, \tag{13}$$



(a) Original 5 cm IMD

(b) Interpolated from 100 cm IMD - linear

(c) Interpolated from 100 cm IMD - proposed

Fig. 5: Estimated direction-of-arrival of direct sound and early reflections for LS 2 (in storage, no continuous line-of-sight between the source and receiver, see Fig. 1a), where IMD denotes inter-measurement distance. Azimuth values are presented from $-170°$ to $190°$ for aided visibility around $\pm180°$, and colour intensity is normalised separately to each measurement's maximum power value.

Table 1: Mean estimated direct sound DoA error E_θ in degrees, between the reference dataset and the test SRIR datasets, where L denotes linear interpolation and P denotes the proposed method. IMD refers to inter-measurement distance of the test SRIR datasets, and LS X refers to the loudspeaker positions as illustrated in Fig. 1a.

IMD (cm)	10	25	50	100	250	500
LS 1 (L)	0.32	0.79	1.08	1.34	2.13	2.69
LS 2 (L)	0.37	0.88	1.68	5.82	25.40	79.70
LS 3 (L)	0.25	0.60	1.42	6.04	20.99	83.00
LS 4 (L)	0.44	0.75	0.88	1.46	3.60	6.30
LS 1 (P)	0.31	0.78	1.08	1.38	2.28	2.74
LS 2 (P)	0.37	0.71	1.29	2.39	3.26	20.87
LS 3 (P)	0.31	0.87	1.37	2.51	4.86	22.46
LS 4 (P)	0.43	0.78	0.87	1.22	3.27	3.33

where $\mathrm{RMS}_{h_j}^{\mathrm{Ref}}$ and $\mathrm{RMS}_{h_j}^{\mathrm{Int}}$ refer to the RMS level of the reference and interpolated SRIRs, respectively, which is calculated for each measurement using (7).

In the vast majority of cases (20/24) the RMS error is greatly reduced with the proposed method: in some cases this is more than a 10x reduction in error. However, there are certain situations where the proposed method produces greater RMS error than the linear method. These are at the greater IMDs, where the acoustical changes between measurements are significant.

Table 2: Mean RMS error E_{RMS} in dB between the reference dataset and the test SRIR datasets. Annotation as in Table 1.

IMD (cm)	10	25	50	100	250	500
LS 1 (L)	0.43	0.59	0.63	0.65	0.52	0.25
LS 2 (L)	0.37	0.52	0.61	0.82	1.50	1.27
LS 3 (L)	0.21	0.34	0.41	0.50	1.36	1.45
LS 4 (L)	0.23	0.35	0.38	0.38	0.33	1.02
LS 1 (P)	0.04	0.04	0.06	0.07	0.07	0.24
LS 2 (P)	0.03	0.08	0.20	0.31	1.54	1.43
LS 3 (P)	0.02	0.08	0.13	0.22	1.30	1.28
LS 4 (P)	0.02	0.03	0.06	0.16	0.60	1.37

2.2 Perceptual Evaluation

To perceptually evaluate the quality of the SRIR interpolation, a listening test was conducted in VR. The test paradigm was a multiple stimulus comparison, with a hidden reference but no anchor. Participants were presented with seven conditions for which they could switch one condition and loudspeaker combination at a time, and were asked to walk the transition and rate the sound quality in terms of overall perceived similarity to the reference, with instructions to listen for all of localisation accuracy, colouration and reverberation. The reference condition was the original dataset of SRIRs, and the test conditions were the interpolated SRIR sets.

Two test stimuli were used: a dry recording of a drumkit, chosen for its transients, sharp attacks, and wide range of frequency content, and an anechoic violin recording, chosen for its smooth and periodic waveform[2]. The SRIRs were auralised using the SPARTA 6DoFconv virtual studio technology (VST) plugin [34][3]. The plugin convolves the SRIRs with a monophonic input signal using fast partitioned time-varying convolution in the frequency domain [35] with the overlap-add method to allow for real-time switching between input SRIRs, with minimal perceptual switching artefacts. The plugin is based on a MATLAB prototype presented in a previous study [11], for which the reader is directed to for a more detailed description. The convolved Ambisonic signals were then rendered binaurally using the SPARTA ambiBIN decoder with Magnitude Least-Squares HRTF pre-processing [36]. Mysphere 3.2 headphones were used for playback, offering high levels of passive transparency [37] which makes them suitable for experiments with both real and virtual sources [11, 12]. Audio processing and programming of the listening test was conducted in Cycling 74 Max.

To display the room transition in VR, three-dimensional models of the two rooms were captured using LiDAR technology from an Apple iPad Pro, with certain features enhanced in post processing, such as the doors and windows, using high resolution two-dimensional textures and sharper edges. Unity was used to render the visuals, which were displayed on an Oculus Quest 2. The loudspeaker model was movable in the environment, such that whichever loudspeaker was currently playing was displayed, as determined in Max and sent to Unity via open sound control (OSC). User position and orientation data, for convolution filter selection and sound field rotation, respectively, was sent from Unity to Max via OSC.

The listening test instructions and multiple stimulus comparison user interface were shown in the Unity virtual environment: the position of these was controlled by the Oculus left hand controller, and interactions made using the trigger on the Oculus right hand controller. To ensure participants stayed within the bounds of the SRIR measurements, a guiding line was placed at 1.2 m above the ground in the Unity scene, from 2.5 m inside the storage space to 2.5 m inside the stairwell, corresponding to the positions of the measurements. In the case that the participant strayed more than 25 cm from the guiding line in the X or Z axis, the screen flashed red and the audio cut out.

The listening test consisted of a total of eight trials: the four loudspeaker positions presented once with the drumkit and once with the violin. No repeats were conducted. Trial and condition ordering was randomised. The tests were conducted on 13 participants aged between 24 and 31 (11 male, 2 female) with self-reported normal hearing and prior critical listening experience (such as education or employment in audio or music engineering).

2.2.1 Results and Discussion

The results of the listening test are presented as violin plots in Fig. 6. Violin plots display both the density trace and box plot, which better illustrates the structure of the data over traditional box plots [38]. The violin widths show the density of data, median values are presented as white points, interquartile ranges are marked using thick grey lines, the ranges between the lower and upper adjacent values are marked using thin grey lines, and individual results are displayed as coloured points.

The results generally show that, with the presented SRIR interpolation method, IMDs up to 50 cm produced perceptually comparable results to the reference at 5 cm IMD. Even for the 100 cm IMD, median values were above 80 for 7 out of 8 tested conditions. At 200 cm and 500 cm IMD, scores were significantly lower, especially for LS 2 and LS 3, where there was no CLOS between the source and receiver, and the largest angular errors in the direct sound direction occur due the choice of direct sound location estimation. This is in fitting with the results shown in [11], which showed that a linear interpolation between the first and last measurements was rated as higher in naturalness for the two sound sources with CLOS (LS 1 and LS 4) than those without (LS 2 and LS 3).

To test the statistical significance of the results, the data was first tested for normality using the Shapiro-Wilk test which showed not all data to be normally distributed, even when excluding the reference condition. Therefore, statistical analysis was conducted using non-parametric methods. Friedman tests showed that the conditions were statistically significantly different ($p < 0.001$) for all stimuli and loudspeaker pairs except LS 1 with the violin stimulus: $\chi^2(6) = 8.32, p = 0.21$; in this configuration both 200 cm and 500 cm IMDs performed relatively well, with median values of 74 and 69, respectively.

The different stimuli, a drumkit and a violin, on the whole produced relatively similar results, though at IMD $\geq$ 100 cm, the median rating of the drumkit was lower for 11 out of 12 cases. This suggests that the drumkit stimulus showed the artefacts of interpolation better, and could suggest that the choice of IMD in measuring could be influenced by the stimuli of the application.

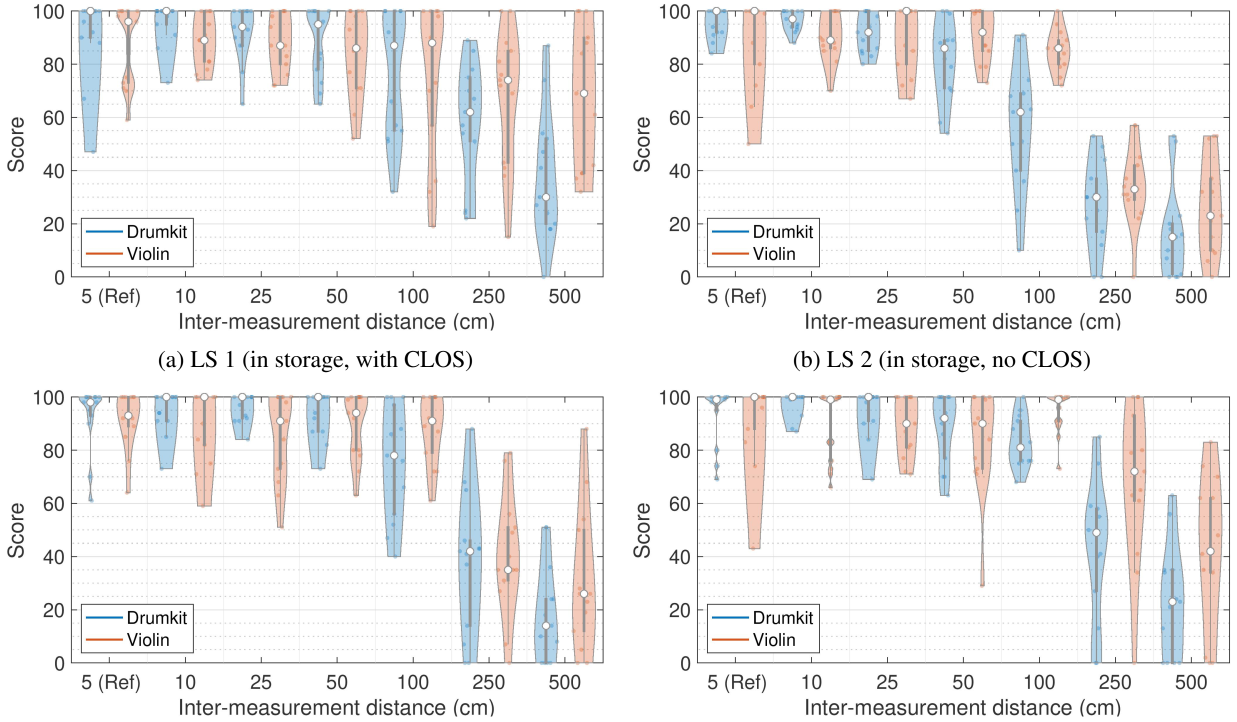
It is notable that there was a broad range of results for each condition, including the reference, which was not rated as 100 by all participants. This is likely a consequence of the listening test being conducted in VR. The cognitive undertaking of performing a multiple stimulus comparison task whereby the auditory experience and conditions are dynamic and dependent on the participant's position and orientation is complex. Multiple thoughts had to be kept in the participant's memory simultaneously and cross-referenced. This is a more complicated task than a standard desktop MUSHRA experience.

3 CONCLUSIONS

This paper has presented a method for auralising measured room transitions in virtual reality. An interpolation

[2] Downloaded from `https://www.openair.hosted.york.ac.uk/`

[3] `https://leomccormack.github.io/sparta-site/docs/plugins/sparta-suite/#6dofconv`



(a) LS 1 (in storage, with CLOS)

(b) LS 2 (in storage, no CLOS)

(c) LS 3 (in stairwell, no CLOS)

(d) LS 4 (in stairwell, with CLOS)

Fig. 6: Violin plots of the perceptual listening test results. CLOS refers to a continuous line-of-sight between the loudspeaker and listener for all listener positions (refer to Fig. 1a for loudspeaker positions and room geometries). Median values are a white point, interquartile range a thick grey line, the range between lower and upper adjacent values a thin grey line, and individual results are coloured points.

method for higher-order spatial room impulse responses (SRIRs) has been proposed, suitable for up to six degrees-of-freedom datasets and robust for interpolating measurements of room transitions. The proposed method uses directional steering of input SRIRs to rotate the direct sounds, equivalent rectangular bandwidth equalisation to directionally equalise early reflections, and RMS compensation throughout to counter level discrepancies. This relatively simple method of interpolation allows for sparser measurement of the acoustics of room transitions whilst retaining a high level of realism in auralisation.

The method has been evaluated numerically, using RMS and direction-of-arrival (DoA) error analysis, which shows that the interpolation is able to reduce error significantly when compared to a linear interpolation method, even up to an inter-measurement distance (IMD) of 100 cm. However, at 250 cm and 500 cm, both methods produce significant errors. A dynamic listening test has then been conducted in virtual reality, using visuals of three-dimensional models from room scans using LIDAR technology and binaural auralisation, where participants were able to walk through the transition in real-time. The conditions of the test were sets of SRIRs generated through interpolation of the original dataset at different IMDs, with the reference the original 5 cm IMD and the lowest case a 500 cm IMD. The results showed that, using the presented interpolation method, IMDs up to 50 cm or in some cases 100 cm were rated as highly similar to the reference.

The results of the evaluation show that, even for a demanding acoustic scenario such as tested in this study, the presented method is able to reduce the necessary inter-measurement distance, which allows for time and cost saving in measurements, with an inter-measurement distance of 50 cm perceptually difficult to discern from a 5 cm reference distance in the tested scenario.

Further work should compare the presented SRIR interpolation method to other available methods. Additionally, the method should be used to interpolate between measurements in a single room, and the results compared to the evaluation in this study, to assess whether a higher IMD is feasible for interpolating between measurements where the acoustical changes are smaller.

The presented interpolation method is available for download as MATLAB code[4], along with demonstration and analysis scripts.

ACKNOWLEDGEMENTS

This research was supported by European Unions Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 812719.

[4] https://github.com/thomas-mckenzie/srir_interpolation

4 REFERENCES

[1] M. A. Gerzon, “Periphony: with-height sound reproduction,” *J. Audio Eng. Soc.*, vol. 21, no. 1, pp. 2–10 (1973).

[2] A. Neidhardt, A. I. Tommy, A. D. Pereppadan, “Plausibility of an interactive approaching motion towards a virtual sound source based on simplified BRIR sets,” presented at the *AES 144th Conv.*, pp. 1–11 (2018 May).

[3] E. Stein, M. M. Goodwin, “Ambisonics depth extensions for six degrees of freedom,” presented at the *AES Int. Conf. on Headphone Technology*, vol. 2019, pp. 1–10 (2019 Aug.).

[4] N. Antonello, E. De Sena, M. Moonen, P. A. Naylor, T. Van Waterschoot, “Room impulse response interpolation using a sparse spatio-temporal representation of the sound field,” *IEEE/ACM Trans. on Audio, Speech and Lang. Proc.*, vol. 25, no. 10, pp. 1929–1941 (2017 Oct.), doi:10.1109/TASLP.2017.2730284.

[5] K. Müller, F. Zotter, “Auralization based on multi-perspective Ambisonic room impulse responses,” *Acta Acustica*, vol. 6, no. 25, pp. 1–18 (2020), doi:10.1051/aacus/2020024.

[6] A. Neidhardt, B. Reif, “Minimum BRIR grid resolution for interactive position changes in dynamic binaural synthesis,” presented at the *AES 148th Conv.*, pp. 1–10 (2020 Jun.).

[7] N. Xiang, Y. Jing, A. C. Bockman, “Investigation of acoustically coupled enclosures using a diffusion-equation model,” *J. Acoust. Soc. Am.*, vol. 126, no. 3, pp. 1187–1198 (2009 Sep.), doi:10.1121/1.3168507.

[8] A. Billon, V. Valeau, A. Sakout, J. Picaut, “On the use of a diffusion model for acoustically coupled rooms,” *J. Acoust. Soc. Am.*, vol. 120, no. 4, pp. 2043–2054 (2006 Oct.), doi:10.1121/1.2338814.

[9] T. McKenzie, S. J. Schlecht, V. Pulkki, “The auditory perceived aperture position of the transition between rooms,” *J. Acoust. Soc. Am.*, vol. 152, no. 3, pp. 1871–1882 (2022 Sep.), doi:10.1121/10.0014178.

[10] T. McKenzie, S. J. Schlecht, V. Pulkki, “Acoustic analysis and dataset of transitions between coupled rooms,” presented at the *IEEE Int. Conf. on Acoust., Speech and Sig. Proc.*, pp. 481–485 (2021 Jun.), doi:10.1109/ICASSP39728.2021.9415122.

[11] T. McKenzie, S. J. Schlecht, V. Pulkki, “Auralisation of the transition between coupled rooms,” presented at the *Int. Conf. on Immersive and 3D Audio*, pp. 1–9 (2021 Sep.), doi:10.1109/I3DA48870.2021.9610955.

[12] N. Meyer-Kahlen, S. Amengual Garí, T. McKenzie, S. J. Schlecht, T. Lokki, “Transfer-plausibility of binaural rendering with different real-world references,” presented at the *Jahrestagung für Akustik - DAGA 2022*, pp. 1–4 (2022 Mar.).

[13] L. McCormack, A. Politis, T. McKenzie, C. Hold, V. Pulkki, “Object-based six-degrees-of-freedom rendering of sound scenes captured with multiple Ambisonic receivers,” *J. Audio Eng. Soc.*, vol. 70, no. 5, pp. 355–372 (2022 May), doi:10.17743/jaes.2022.0010.

[14] Axel Plinge, S. J. Schlecht, O. Thiergart, T. Robotham, O. Rummukainen, E. A. P. Habets, “Six-degrees-of-freedom binaural audio reproduction of first-order Ambisonics with distance information,” presented at the *AES Conf. on Audio for Virtual and Augmented Reality* (2018 Aug.), doi:10.1016/s0967-2109(98)00021-0.

[15] J. G. Tylka, E. Y. Choueiri, “Domains of practical applicability for parametric interpolation methods for virtual sound field navigation,” *J. Audio Eng. Soc.*, vol. 67, no. 11, pp. 882–893 (2019 Nov.), doi:10.17743/JAES.2019.0038.

[16] C. Masterson, G. Kearney, F. Boland, “Acoustic impulse response interpolation for multichannel systems using Dynamic Time Warping,” presented at the *AES 35th Int. Conf.*, pp. 1–10 (2009 Feb.).

[17] O. Das, P. Calamia, S. V. A. Gari, “Room impulse response interpolation from a sparse set of measurements using a modal architecture,” presented at the *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 960–964 (2021 jun).

[18] J. Zhao, X. Zheng, C. Ritz, D. Jang, “Interpolating the directional room impulse response for dynamic spatial audio reproduction,” *Applied Sciences*, vol. 12, no. 4 (2022 Feb.), doi:10.3390/app12042061.

[19] E. Fernandez-Grande, D. Caviedes-Nozal, M. Hahmann, X. Karakonstantis, S. A. Verburg, “Reconstruction of room impulse responses over extended domains for navigable sound field reproduction,” presented at the *Int. Conf. on Immersive and 3D Audio* (2021 Sep.), doi:10.1109/I3DA48870.2021.9610846.

[20] G. Götz, S. J. Schlecht, V. Pulkki, “Common-slope modeling of late reverberation,” *TechRxiv. Preprint*, pp. 0–13 (2022), doi:10.36227/techrxiv.20482767.v1.

[21] B. C. Moore, B. R. Glasberg, “Suggested formulae for calculating auditory-filter bandwidths and excitation patterns,” *J. Acoust. Soc. Am.*, vol. 74, no. 3, pp. 750–753 (1983 Sep.), doi:10.1121/1.389861.

[22] A. Lindau, L. Kosanke, S. Weinzierl, “Perceptual evaluation of model- and signal-based predictors of the mixing time in binaural room impulse responses,” *J. Audio Eng. Soc.*, vol. 60, no. 11, pp. 887–898 (2012 Nov.).

[23] T. Hidaka, Y. Yamada, T. Nakagawa, “A new definition of boundary point between early reflections and late reverberation in room impulse responses,” *J. Acoust. Soc. Am.*, vol. 122, no. 326 (2007 Jul.), doi:10.1121/1.2743161.

[24] M. R. Schroeder, “New method of measuring reverberation time,” *J. Acoust. Soc. Am.*, vol. 37, no. 6, pp. 1187–1188 (1965 Jun.), doi:10.1121/1.1939454.

[25] K. Meesawat, D. Hammershøi, “An investigation on the transition from early reflections to a reverberation tail in a BRIR,” presented at the *Int. Conf. on Auditory Display*, pp. 5–9 (2002 Jul.).

[26] A. Campos, S. Sakamoto, C. D. Salvador, “Directional early-to-late energy ratios to quantify clarity: A case study in a large auditorium,” presented at the *Int. Conf. on Immersive and 3D Audio* (2021 Sep.).

[27] C. Hold, S. J. Schlecht, A. Politis, V. Pulkki, “Spatial filter bank in the spherical harmonic domain: reconstruction and application,” presented at the *IEEE*

Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 361–365 (2021 Oct.), doi:10.1109/WASPAA52581.2021.9632709.

[28] C. Hold, T. McKenzie, G. Götz, S. J. Schlecht, V. Pulkki, "Resynthesis of spatial room impulse response tails with anisotropic multi-slope decays," *J. Audio Eng. Soc.*, vol. 70, no. 6, pp. 526–538 (2022 Jun.), doi:10.17743/jaes.2022.0017.

[29] R. H. Hardin, N. J. Sloane, "McLaren's improved snub cube and other new spherical designs in three dimensions," *Discrete Comp. Geom.*, vol. 15, pp. 429–441 (1996), doi:10.1109/isit.1995.531530.

[30] P. Majdak, A. E. S. Member, F. Zotter, A. E. S. A. Member, "Spatially Oriented Format for Acoustics 2.1: introduction and recent advances," *J. Audio Eng. Soc.*, vol. 70, no. 7, pp. 565–584 (2022 Jul./Aug.).

[31] G. Götz, C. Hold, T. McKenzie, S. J. Schlecht, V. Pulkki, "Analysis of multi-exponential and anisotropic sound energy decay," presented at the *Jahrestagung für Akustik - DAGA 2022*, pp. 1–4 (2022 Mar.).

[32] J. Daniel, S. Moreau, "Further study of sound field coding with higher order Ambisonics," presented at the *AES 116th Conv.*, pp. 1–14 (2004 May).

[33] A. Politis, *Microphone array processing for parametric spatial audio techniques*, Phd thesis, Aalto University (2016).

[34] T. McKenzie, N. Meyer-kahlen, R. Daugintis, L. Mccormack, S. J. Schlecht, V. Pulkki, "Perceptually informed interpolation and rendering of spatial room impulse responses for room transitions," presented at the *Int. Congress on Acoustics*, pp. 1–11 (2022 Oct.).

[35] F. Wefers, M. Vorländer, "Efficient time-varying FIR filtering using crossfading implemented in the DFT domain," presented at the *Forum Acusticum* (2014 Sep.).

[36] C. Schörkhuber, M. Zaunschirm, R. Höldrich, "Binaural rendering of Ambisonic signals via magnitude least squares," presented at the *Fortschritte der Akustik – DAGA 2018*, pp. 339–342 (2018 Mar).

[37] P. Llado, T. McKenzie, N. Meyer-Kahlen, S. J. Schlecht, "Predicting perceptual transparency of head-worn devices," *J. Audio Eng. Soc.*, vol. 70, no. 7/8, pp. 585–600 (2022 Jul./Aug.), doi:10.17743/jaes.2022.0024.

[38] J. L. Hintze, R. D. Nelson, "Violin plots: a box plot-density trace synergism," *Am. Statistician*, vol. 52, no. 2, pp. 181–184 (1998 May).

THE AUTHORS

Thomas McKenzie

Nils Meyer-Kahlen

Christoph Hold

Sebastian J. Schlecht

Ville Pulkki

Thomas McKenzie is a Lecturer in Acoustics and Architectural Acoustics at the Reid School of Music, Edinburgh College of Art, University of Edinburgh, UK. He completed a BSc in Music, Multimedia and Electronics at the University of Leeds, UK in 2013, before completing his MSc in Postproduction with Sound Design and then PhD in Music Technology at the University of York, UK in 2015 and 2020, respectively. He then undertook a position as a postdoctoral researcher in the Department of Signal Processing and Acoustics at Aalto University, Finland, where he studied room acoustics and six degrees-of-freedom spatial audio, as part of the Business Finland funded Human Optimised XR project. His research interests include spatial audio and psychoacoustics.

Nils Meyer-Kahlen is a doctoral candidate for the Department of Signal Processing and Acoustics at Aalto University, Finland. Before joining the lab in 2019, he studied Electrical Engineering and Audio Engineering at the Technical University and the University of Music and Performing Arts in Graz, Austria. His main interests are the technology and perception of spatial sound. Currently, he studies room acoustic perception in virtual and augmented realities.

Christoph Hold is a doctoral candidate in the Department of Signal Processing and Acoustics at Aalto University, Finland, focusing on spatial audio processing. He received a M.Sc. in audio communication and technology in 2019 and a B.Sc. in electrical engineering from the Technische Universität Berlin, where he specialized in signal processing and virtual acoustics. From 2015 to 2017 he was a research assistant at TU Berlin, followed by two research internships (2017 and 2018) at Microsoft Research in Redmond, WA, USA. He is interested in high quality audio and its perception. For the Audio Engineering Society, he was the chair of the Berlin Student Section and part of the 142nd AES Convention committee.

Sebastian J. Schlecht is a Professor of Practice for Sound in Virtual Reality at the Acoustics Lab, Department of Signal Processing and Acoustics and Media Labs, Department of Media, of Aalto University, Finland. He received the Diploma in Applied Mathematics from the University of Trier, Germany, in 2010 and an MSc degree in Digital Music Processing from the School of Electronic Engineering and Computer Science at Queen Mary University of London, UK, in 2011. In 2017, he received a Doctoral degree at the International Audio Laboratories Erlangen, Germany, on artificial spatial reverberation and reverberation enhancement systems. From 2012 to 2019, Dr. Schlecht was also an external research and development consultant and lead developer of the 3D Reverb algorithm at the Fraunhofer IIS, Erlangen, Germany.

Ville Pulkki is a professor in the Department of Signal Processing and Acoustics at Aalto University, Helsinki, Finland. He has been working in the field of spatial audio for over 20 years. He developed the vector-base amplitude panning (VBAP) method in his Ph.D. (2001) and directional audio coding after the Ph.D. with his research group. He also has contributions in perception of spatial sound, laser-based measurement of room responses, and binaural auditory models. He has received the Samuel L. Warner Memorial Medal Award from SMPTE and the AES Silver Medal Award. He enjoys being with his family, building his summer house, and performing in musical ensembles.