OPEN ACCESS Freely available online

PAPERS

L. McCormack, V. Pulkki, A. Politis, O. Scheuregger and M. Marschall "Higher-Order Spatial Impulse Response Rendering: Investigating the Perceived Effects of Spherical Order, Dedicated Diffuse Rendering, and Frequency Resolution"
J. Audio Eng. Soc., vol. 68, no. 5, pp. 338–354, (2020 May.).
DOI: https://doi.org/10.17743/jaes.2020.0026

Higher-Order Spatial Impulse Response Rendering: Investigating the Perceived Effects of Spherical Order, Dedicated Diffuse Rendering, and Frequency Resolution

LEO MCCORMACK[1], *AES Student Member*, **VILLE PULKKI**[1,2], *AES Fellow*, **ARCHONTIS POLITIS**[1,3], *AES Member*, **OLIVER SCHEUREGGER**[2], **AND MARTON MARSCHALL**[2], *AES Member*

[1]*Aalto University, Espoo, Finland*
[2]*Technical University of Denmark, Kongens Lyngby, Denmark*
[3]*Tampere University, Tampere, Finland*

This article details an investigation into the perceptual effects of different rendering strategies when synthesizing loudspeaker array room impulse responses (RIRs) using microphone array RIRs in a parametric fashion. The aim of this rendering task is to faithfully reproduce the spatial characteristics of a captured space, encoded within the input microphone array RIR (or the spherical harmonic RIR derived from it), over a loudspeaker array. For this study, a higher-order formulation of the Spatial Impulse Response Rendering (SIRR) method is introduced and subsequently employed to investigate the perceptual effects of the following rendering configurations: the spherical harmonic input order, frequency resolution, and utilizing dedicated diffuse stream rendering. Formal listening tests were conducted using a 64-channel loudspeaker array in an anechoic chamber, where simulated reference scenarios were compared against the outputs of different methods and rendering configurations. The test results indicate that dedicated diffuse stream rendering and higher analysis orders both yield noticeable perceptual improvements, particularly when employing problematic transient stimuli as input. Additionally, it was found that the frequency resolution employed during rendering has only a minor influence over the perceived accuracy of the reproduction in comparison to the other two tested attributes.

0 INTRODUCTION

Over the years, a handful of methods have been proposed for the task of reproducing spatial room impulse responses (RIRs) over loudspeaker arrays. Essentially, these methods aim to solve the problem of generating suitable loudspeaker RIRs using microphone array RIRs, such that (after convolution) an anechoic signal may be reproduced and exhibit all of the spatial characteristics of the captured space. Applications of such methods include the following: auralizing existing acoustical spaces for aesthetic purposes, artistic production, architectural design, and psychoacoustic studies regarding the auditory perception of spaces. Existing methods for this task may be loosely categorized as being either nonparametric or parametric based. The former class of methods rely on a linear and time-invariant mapping of the microphone array RIR to the loudspeaker RIR, whereas parametric methods often employ prior knowledge of the structure of RIRs and take the resolution of human spatial hearing into consideration. Essentially, these parametric alternatives operate in two stages: by first estimating perceptually meaningful parameters to describe the captured response and subsequently employing those parameters to conduct the mapping of the microphone array RIR to the loudspeaker array RIR in a more informed manner.

This work focuses primarily on parametric methods which employ spherical harmonic (also referred to as Ambisonic or B-Format) RIRs as input and loudspeaker RIRs as output; however, binaural RIRs may also be subsequently derived from these loudspeaker RIRs. Methods formulated

in the spherical harmonic domain (SHD) are of particular interest, as the microphone array specifications are largely abstracted away from the algorithms that employ them. This permits the use of arbitrary 3D microphone arrays, provided suitable spatial encoding is first conducted [1]. This important trait bestows a degree of flexibility not found with algorithms that rely on specific types of microphone arrays with known specifications. Furthermore, spherical harmonic components are trivial to synthesize, and many commercially available spherical microphone arrays have been designed specifically for the task of obtaining them. Note that spherical arrays are generally the favored array format for spherical harmonic acquisition due to their uniform full spherical "field of view" (capturing the reflections and reverberation equally from all directions) and are therefore particularly prevalent in the field of room acoustics research.

0.1 Motivation for This Study

There are a vast number of possibilities for reproducing spatial RIRs in a parametric fashion. However, there has not yet been a systematic evaluation of the perceived effects of different design choices. The authors of the Spatial Decomposition Method (SDM) [2] claim that it is able to produce authentic reproductions by employing broad-band direction of arrival (DoA) analysis in narrow temporal windows and assigning the pressure signal to the loudspeaker setup accordingly, whereas the Spatial Impulse Response Rendering (SIRR) method [3, 4] estimates the DoA in frequency bands and conducts separate processing for the diffuse and nondiffuse components that comprise the input sound-field. Therefore, the first aim of this study was to quantify the effects of these design choices through formal listening tests, namely the effects of increasing **frequency resolution** and the effect of **dedicated diffuse stream rendering**.

Furthermore, since higher-order parametric processing has been shown to be advantageous for complex acoustic scenarios when using spherical harmonic *signals* as input [5–7], it is assumed that similar higher-order processing would also be beneficial for spatial RIRs. Therefore, a new higher-order formulation of the SIRR method is also presented in this article; and thus, an additional aim of this study was to ascertain the perceived effects of increasing the **spherical harmonic input order** when using this expanded formulation.

0.2 Organization of the Article

This article begins with background information regarding existing methods for generating loudspeaker RIRs using spherical harmonic RIRs as input. Sec. 2 describes the reformulated SIRR method, which can accommodate higher-order spherical harmonic RIRs as input and features an anisotropic diffuse stream rendering approach. The listening test environment and room simulation tools, used to synthesize the reference scenarios, are then described in Sec. 3. This is followed by informal observations in Sec. 4, regarding the perceived differences between the different parametric rendering strategies. Formal listening tests are then described in Sec. 5, where the effects of the following were investigated: increasing the spherical harmonic input order, increasing the frequency resolution, and the use of dedicated diffuse stream rendering. The reformulated SIRR method (with different rendering configurations) and the first-order spherical harmonic variant of SDM [8] were employed for this task. Discussions pertaining to these formal listening test results are then presented in Sec. 6, and the article is concluded in Sec. 7.

1 BACKGROUND

In the most basic scenario in communication acoustics, a directional source emits sound into a space, and this sound subsequently arrives at the listening position via multiple travelling pathways. Upon studying the room impulse response (RIR) of this acoustical chain, a number of peaks are typically observable in the early part of the response, with each peak corresponding to an individual reflection from surfaces within the space. Over time, the number of reflections arriving at the listening position grows exponentially, and many succumb to diffraction effects caused by edges and diffusion effects caused by rough surfaces. This late part of the impulse response often resembles, and may also be modeled in practice as, exponentially decaying noise [9]. As these sound components arrive at the listening position, the listener is typically able to perceive certain spatial attributes related to their surroundings, namely the direction of first arrival, the directions of prominent reflections (not fused by the precedence effect), the distance of the source, and the directional energy distribution of reverberation [10–12].

However, in order to capture the spatial attributes of a space, the sound-field pressure must be sampled at multiple positions. In practice, a microphone array is employed for this task, and thus, the necessary spatial information is encapsulated within the microphone array RIR or the spherical harmonic RIR derived from it. A reproduction method is then responsible for synthesizing a suitable loudspeaker RIR using this input. The primary design philosophy is arranged so that, after convolving with this loudspeaker RIR, an anechoic source may be reproduced over the target loudspeaker array and exhibit all of the spatial attributes of the measured space. In the ideal case, the original 3D sound-field captured in the measured space should be exactly reproduced at the listening position. However, due to practical limitations, a perceptually authentic rendition of the captured response is often the most realistic goal of any reproduction method. Naturally, the listening room should also be as anechoic as possible, or the reproduced response should be experienced over headphones, in order to avoid conflicting spatial cues between the reproduced space and the physical space surrounding the listener. Note that during reproduction, the convolved stimuli will also exhibit the same directivity characteristics of the excitation source used to measure the spatial RIR.

As mentioned earlier, there are two subclasses of methods which may be employed for this reproduction task, described as being either nonparametric or parametric based. Both classes of methods are described in more detail below.

1.1 Nonparametric Reproduction Methods

Nonparametric methods aim to replicate the original pressure field, at all audible frequencies, using a linear combination of the input RIRs. However, this task is especially challenging from both a technical and physical point of view, as the human auditory system is sensitive to a vast range of wavelengths, from approximately 30 meters to 2 centimeters [10]. Ambisonics [13] is one popular example of a method developed with this starting point. However, concurrently reproducing all audible frequencies with high spatial accuracy has been shown to be impossible in practice [14–17].

The main appeal of nonparametric methods is that they have low implementation complexity and computational requirements when compared to their parametric counterparts. They also do not introduce any distortions or time-varying artifacts into the output. However, nonparametric methods are also inherently limited by the spatial resolution of the input format, which is largely dictated by the number of microphones (or the order of spherical harmonic components). For example, first and lower-order Ambisonics has been found to exhibit perceptual deficiencies such as the directional blurring of point sources, localization ambiguity, reduced sense of envelopment in reverberant conditions, and strong coloration effects (including comb-filtering) [18, 19, 15, 20, 17]. These deficiencies are largely due to the high signal coherence between loudspeaker channels, as it is not possible to generate spatial patterns of sufficiently narrow beam-widths at these low input orders.

1.2 Parametric Reproduction Methods

As previously discussed, typically, RIRs exhibit a clear structure. The motivation for using parametric techniques is the notion that different components of the response may be more appropriately reproduced by employing dedicated strategies, which are more targeted towards them. Therefore, parametric methods operate by first identifying these different components and extracting suitable parameters to describe their behavior. For example, the most commonly extracted components and associated parameters are the first-arriving components and their corresponding direction-of-arrival (DoA) estimates, which relate to the pressure peaks in the early response. Since these peaks are generally caused by distinct reflections, perhaps an intuitively reasonable course of action is to assign them directly to the loudspeaker setup (based on the estimated DoAs) in a synthesis stage. This would ultimately result in a higher output spatial resolution and a sharper perceived image than is otherwise possible when employing linear alternatives with the same order of input.

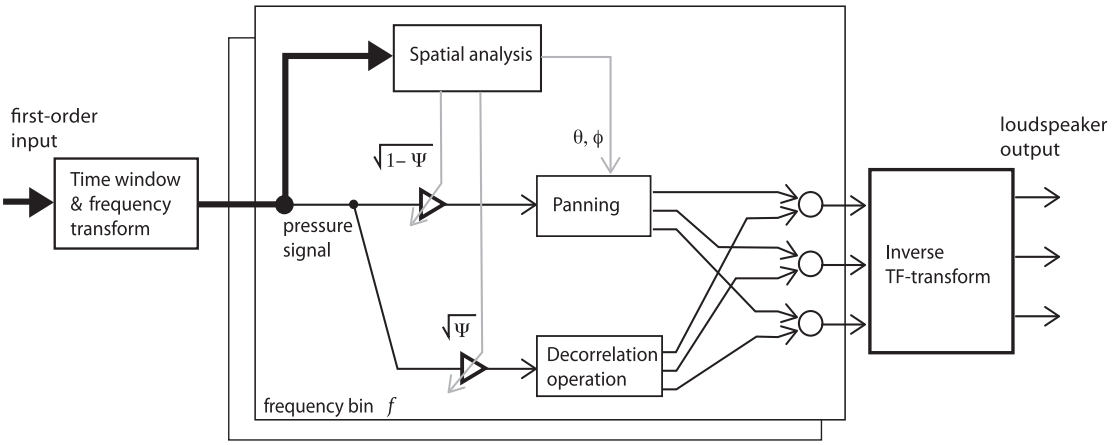
The components a parametric method deems to be important is dictated by its sound-field model. A sound-field model is essentially a set of assumptions, which are made regarding the composition and behavior of the sound-field. Therefore, there are two key challenges when designing a parametric method. The first is to identify an appropriate sound-field model, which describes the input scene in a perceptually and/or physically meaningful manner, while the second is to employ the appropriate signal processing techniques so as to robustly synthesize the output in a manner that not only conforms to the chosen sound-field model, but also does not incur perceivable artifacts in the process.

There exists only a few sound-field models and synthesis approaches, which have been proposed specifically for the task of rendering spatial RIRs [3, 4, 21, 2, 22–26]. Perhaps the most popular are the two methods employed for the formal listening tests in Sec. 5, namely SIRR [3, 4] and SDM [2, 8], which are described in detail in the following text.

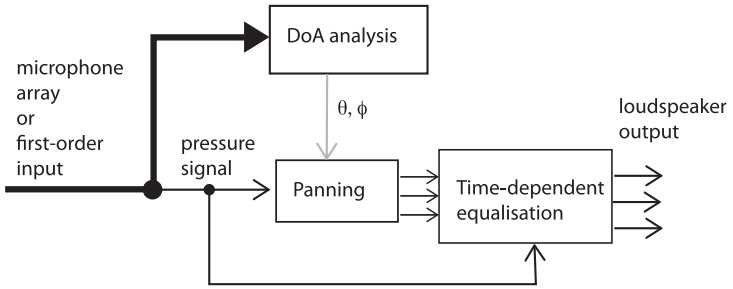
1.2.1 Spatial Impulse Response Rendering

The SIRR [3, 4] method was the first parametric reproduction method proposed for spatial RIRs, which built upon the spatial analysis techniques described in [27, 28–30]; a flow diagram of the method is depicted in Fig. 1(a). The method employs a sound-field model that assumes the existence of a single time-varying directional component per narrow-band frequency and an isotropic diffuse-field. In practice, the method operates in the time-frequency domain and employs a first-order spherical harmonic RIR as input and comprises dedicated analysis and synthesis stages. In the analysis stage, spatial parameters are extracted for each time-frequency tile, often via the energetic properties of the active-intensity vector [27, 31]. Since the active-intensity vector points in the direction of the flow of acoustical energy, an assumption is made that the opposite direction corresponds to the DoA of a source signal. The diffuseness parameter may then be derived based on the ratio of the intensity vector length and the total sound-field energy [32]. However, note that when using a real microphone array, the spherical harmonic components cannot be derived at all frequencies. Therefore, provided that the geometry of the microphone array is known, space-domain alternatives may be employed for the parameter extraction at these frequencies.

In the synthesis stage, the output loudspeaker RIR is generated by panning the omnidirectional (pressure) component to the analyzed DoA using Vector-Base Amplitude-Panning (VBAP) [33] and also replicating this omnidirectional component to all loudspeaker channels—which are then subjected to decorrelation operations. Note that the intention of this decorrelation is to *scramble* the phase response for each loudspeaker channel, while leaving the amplitude response intact. The balance between these nondiffuse and diffuse streams is then dictated by the temporal and frequency-dependent diffuseness parameter, which has a value between zero (fully nondiffuse) and one (fully diffuse). Ideally, this processing serves to route the time-domain peaks, caused by the direct sound and distinct reflections, to the nondiffuse stream. These peaks are then reproduced as point-like phantom sources corresponding



(a) Spatial impulse response rendering (SIRR)



(b) Spatial decomposition method (SDM)

Fig. 1. Flow diagrams of the SIRR and SDM architectures. Note that SIRR conducts the spatial analysis and synthesis for each frequency bin independently. The DoA analysis for the SDM method may be conducted via the cross-correlations between spaced omnidirectional sensors [27, 2] or using the intensity vector derived from first-order spherical harmonic input in the same manner as SIRR [8]. The time-dependent equalization is then applied as described in [35].

to the estimated source/reflection directions. The diffuse stream should then comprise only the components which contribute to the reverberant diffuse-field and have a temporally smooth envelope.

Since the main source of problems for the Ambisonics method is the erroneously high coherence between loudspeaker channels, it is evident that the SIRR processing, for both diffuse and nondiffuse streams, mitigates this problem. Therefore, loudspeaker RIRs rendered by SIRR tend to exhibit reduced coloration problems and improved spatial accuracy, when compared with Ambisonics of the same order; this is also shown to be the case in Sec. 5. However, when rendering a loudspeaker RIR in such a manner, there are a number of potential issues that may occur. For example, the diffuseness value may deviate from the extremities in certain cases (i.e., it may lie somewhere between zero and one), which may arise when a pressure peak of a reflection is accompanied by a considerable amount of diffuse energy. Here, the SIRR algorithm is unable to distinguish between the reflection and diffuse components, and they are thus routed to both the diffuse and nondiffuse streams.

Another limitation of the method is that its model assumes the presence of a single reflection (or no reflections) for each temporal analysis window and frequency bin. Therefore, in cases in which two or more reflections arrive simultaneously, the diffuseness will be estimated to be higher than in reality, resulting in some direct components being erroneously routed to the diffuse stream and subsequently decorrelated. This would lead to the temporal smearing of some reflections, which may be audible to the listener. Furthermore, the original SIRR formulation also forced the diffuse stream to be isotropic, which may not correspond to that of the original space. This latter limitation may, however, be minimized by employing cardioid beampatterns for each loudspeaker direction (akin to ambisonic decoding), rather than replicating the omnidirectional component for each loudspeaker channel, as was suggested later in [32]. However, the diffuse stream is still modulated with direction-independent diffuseness values.

1.2.2 Spatial Decomposition Method

The SDM [2] is based on the assumption that a measured pressure pulse, at a given point in space, is the result of a single broad-band image source varying rapidly in direction and amplitude over time. It was developed to overcome SIRR input limitations by employing the microphone array RIRs directly, since, while the SHD is often

considered a more convenient format, some loss of spatial performance is incurred during the conversion. The main flow diagram of the SDM architecture is depicted in Fig. 1(b). The original formulation of the method operated using open spherical microphone arrangements of four or more omnidirectional sensors. The microphone array RIR is then analyzed using short time windows, with the DoA estimated via the cross-correlations between the microphone channels, as proposed earlier in [27]. The authors of SDM also later developed a popular first-order SHD variant (SDM-B-Format) [8], which estimates the DoA based on a broad-band intensity vector in the same manner as SIRR, albeit without frequency resolution or dedicated diffuse stream rendering. During the synthesis stage, the pressure signal is either quantized to the nearest loudspeaker or panned using VBAP. Since SDM was originally intended for concert hall auralization, in [34], it was argued that quantization is more preferable as timbral colorations due to panning are mitigated.

It is important to note that SDM does not explicitly divide the input sound-field into separate diffuse and nondiffuse streams. While it is intuitively clear that the peaks in the early response (which correspond to the direct sound and early reflections) will likely be reproduced as intended, it is less clear what happens when reverberation is introduced into the later part of the response. Here, the assumption made with the SDM model is that the DoA will vary randomly, subsequently distributing the diffuse energy equally in all directions, and thus produce an evenly distributed reverberant tail. However, it is a concern that this implicit assumption may not always hold in practice. Another concern with SDM arises due to the relatively short analysis windows, approximately 1 ms, which corresponds to the period of a 1-kHz sinusoid. This essentially means that at lower frequencies, each cycle of the waveform is divided into shorter temporal components and subsequently reproduced based on directions defined by broad-band DoA estimates. Therefore, this processing may cause a perceivable degree of distortion, which, in turn, alters the spectral balance of the response. Indeed, the authors of SDM did identify and attempt to mitigate this latter problem through an adaptive post-equalization operation proposed in [35], which aims to match the mean of the loudspeaker magnitude responses with that of the input pressure magnitude response.

In [2], the reproduction quality of SDM was studied through formal listening tests with pair-wise comparisons between reference scenarios and reproductions using both SDM and SIRR. Reference scenarios with different reverberation times were utilized, and the input stimuli were speech, trombone, and castanets. The SIRR implementation, by the authors of SDM, produced large discrepancies between the reference scenario and the reproduction, whereas SDM yielded a prominently more authentic reproduction. For the listening test, the authors simulated a six-sensor array for the SDM test cases and applied additional spatial encoding to obtain first-order spherical harmonic components for the SIRR test cases. However, it should be noted that open arrays with omnidirectional sensors are not well suited to this conversion [36], and the authors also did not detail how they conducted this conversion. Furthermore, SIRR and SDM-B-Format implementations developed by the respective original authors have been employed for the comparisons in Sec. 4 and 5, which show vastly different results to those presented in [2].

2 SPATIAL IMPULSE RESPONSE RENDERING WITH HIGHER-ORDER SPHERICAL HARMONIC INPUT

In this section, a new higher-order SIRR (HO-SIRR) architecture is presented[1]. The principles behind processing higher-order spherical harmonic input have been previously derived for the Directional Audio Coding (DirAC) method [6, 38], which is intended for reproducing *continuous* spherical harmonic signals. Its primary design philosophy relates to the idea that (provided higher-order input is available) the sound-field may be first segregated into individual sectors, which comprise spatially localized pressure and pressure gradient components. These spatially localized variants of the pressure and pressure gradients form the basis for estimating an intensity-vector, which is biased towards favoring energetic contributions to the sound-field in one region on the sphere. Therefore, DoA and diffuseness estimates made in one sector, will have reduced susceptibility to noise and interferers present in other sectors. Additionally, the degree to which the sound-field may be segregated is directly proportional to the input order, which renders the approach quite scalable with increasing input resolution. Note that a flow diagram of the HO-SIRR architecture is depicted in Fig. 2.

By dividing the sound-field into sectors, problematic cases, such as multiple reflections arriving simultaneously from different directions, are analyzed more reliably [39] (assuming that these individual reflections do indeed fall within their own sector). Therefore, panning the spatially localized pressure components based on local DoA estimates, rather than panning the whole sound-field pressure based on a global DoA estimate, represents a more robust approach to rendering the direct stream when compared to first-order SIRR.

For the diffuse-stream, the sector signals are first scaled according to their respective diffuseness estimates and re-encoded back into the SHD. This SHD representation of the diffuse-stream is then distributed to the loudspeakers using an Ambisonic decoder, followed by decorrelation of each loudspeaker channel. Note that these diffuseness estimates are direction-dependent, since the sectors are steered in different directions on the sphere. Therefore, this may also allow the method to more faithfully reproduce anisotropic diffuse-fields, since each sector direction acts as a control point on the sphere for manipulating the amount of diffuse energy. Furthermore, although the diffuseness value estimated for each sector is generally lower than that of the original sound-field, the reproduced field still exhibits the

[1] Note that a preliminary formulation of this higher-order architecture was first detailed in [37].

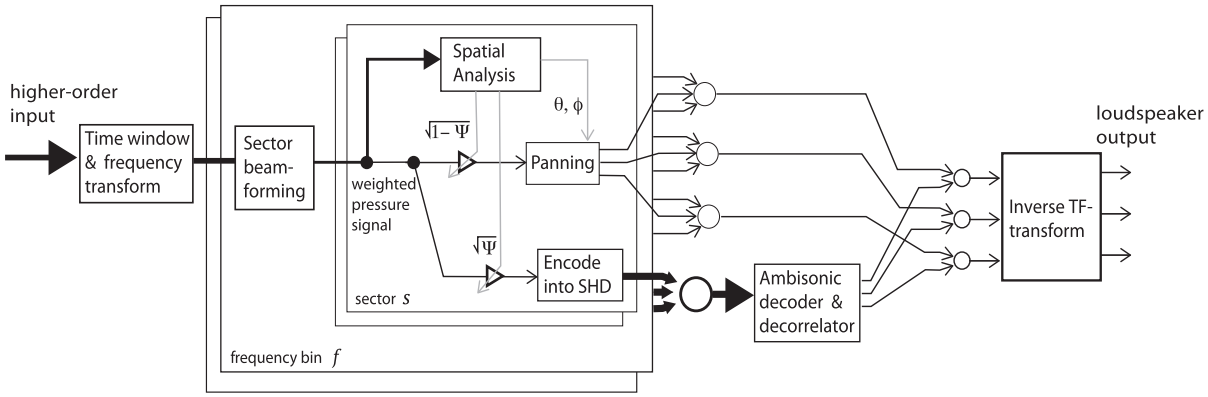


Fig. 2. Flow diagram of the HO-SIRR architecture. The analysis is conducted for each frequency bin and sector, independently. The nondiffuse and diffuse components are then summed over all sectors for each frequency bin. The resulting SHD diffuse stream is distributed to the loudspeakers using an Ambisonics decoder, followed by decorrelation.

same degree of diffuseness as when estimating the diffuseness for the whole (nonsegregated) sound-field [40]. Therefore, these inherently lower diffuseness values subsequently lead to the reduction of artifacts caused by decorrelation, as the energy of the diffuse stream is reduced.

2.1 Ambisonic Decoding

Perhaps the most common linear reproduction method, based on spherical harmonic input, is Ambisonics [13]. In the HO-SIRR formulation, Ambisonics is employed for the task of redistributing the diffuse components to the loudspeaker channels. The Nth order Ambisonic decoding of spherical harmonic RIRs, $\mathbf{a} \in \mathbb{R}^{(N+1)^2 \times 1}$, to L loudspeaker channels may be conducted via the application of a single decoding matrix of real-valued gains $\mathbf{D}_{\text{ls}} \in \mathbb{R}^{L \times (N+1)^2}$. The loudspeaker RIRs, $\mathbf{y} \in \mathbb{R}^{L \times 1}$, are thus obtained as

$$\mathbf{y}_{\text{ambi}}(t) = \mathbf{D}_{\text{ls}}\mathbf{a}(t). \tag{1}$$

There have been many proposed solutions for determining the decoding matrix values [41, 42], including

Sampling:	$\mathbf{D}_{\text{ls}} = \frac{1}{L}\mathbf{Y}_{\text{ls}}^{\text{T}}$
Mode-Matching:	$\mathbf{D}_{\text{ls}} = (\mathbf{Y}_{\text{ls}}^{\text{T}}\mathbf{Y}_{\text{ls}})^{-1}\mathbf{Y}_{\text{ls}}^{\text{T}}$
AllRAD:	$\mathbf{D}_{\text{ls}} = \frac{1}{K}\mathbf{G}_{\text{td}}\mathbf{Y}_{\text{td}}^{\text{T}}$

where $\mathbf{Y}_{\text{ls}} \in \mathbb{R}^{(N+1)^2 \times L}$ is a matrix of spherical harmonic weights for each loudspeaker direction; $\mathbf{Y}_{\text{td}} \in \mathbb{R}^{(N+1)^2 \times K}$ are the spherical harmonic weights for K directions of a uniformly distributed arrangement of points on the sphere (e.g., a *t-design* [43]); and $\mathbf{G}_{\text{td}} \in \mathbb{R}^{L \times K}$ are VBAP [33] gains for panning the uniformly arranged *virtual loudspeakers* to the target loudspeaker setup.

2.2 Legacy Spatial Impulse Response Rendering

The HO-SIRR formulation is based on the original SIRR method [3, 4], which operated on first-order, $N = 1$, spherical harmonic RIRs
$\mathbf{a_1}(t, f) = [\text{a}_{00}(t, f), \text{a}_{1(-1)}(t, f), \text{a}_{10}(t, f), \text{a}_{11}(t, f)]^{\text{T}}$, in the time-frequency domain, where t and f refer to the time and frequency indices, respectively. A parameter vector, $\mathbf{p}_1 = \mathcal{A}[\mathbf{a_1}] = [\theta, \phi, \psi]$, is then estimated for each time-frequency tile, which comprises the azimuth-elevation angles (θ, ϕ) and the diffuseness (ψ). As suggested in the original publication [3], these parameters may be extracted via the energetic properties of the active-intensity vector [27, 31], which may be derived from the zeroth and first-order spherical harmonic components (note that the time and frequency indices are henceforth omitted for the brevity of notation)

$$\mathbf{i}_{\text{a}} = \Re[p\mathbf{u}^*], \tag{2}$$

with[2]

$$p \simeq \text{a}_{00}, \text{ and } \mathbf{u} \simeq -\frac{1}{\rho_0 c\sqrt{3}}\begin{bmatrix} \text{a}_{11} \\ \text{a}_{1(-1)} \\ \text{a}_{10} \end{bmatrix}, \tag{3}$$

where $\Re$ denotes the real operator, * denotes the complex conjugate operator, p is the sound pressure, ρ_0 is the mean density of the medium, c is the speed of sound, and $\mathbf{u}$ is the particle velocity. Note that the particle velocity may be used in this manner, with the assumption that the sound sources are in the far-field [44]. The parameters may then be estimated as

$$\theta, \phi = \angle - \frac{\mathbf{i}_{\text{a}}}{||\mathbf{i}_{\text{a}}||}, \text{ and } \psi = 1 - \frac{2||\mathbf{i}_{\text{a}}||}{|p|^2 + \mathbf{u}^{\text{H}}\mathbf{u}}. \tag{4}$$

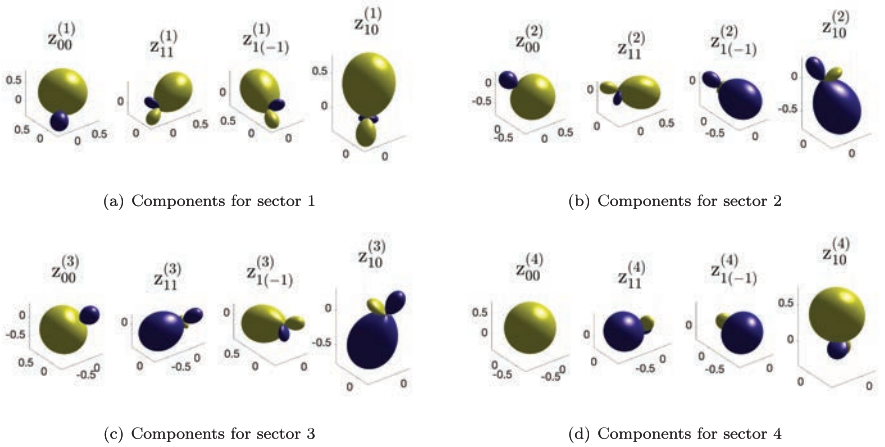
The direct $(\mathbf{y}_{\text{dir}} \in \mathbb{R}^{L \times 1})$ and diffuse $(\mathbf{y}_{\text{diff}} \in \mathbb{R}^{L \times 1})$ loudspeaker RIRs are then independently synthesized by employing the analyzed parameters as

$$\mathbf{y}_{\text{dir}} = \sqrt{1 - \psi}\mathbf{g}(\theta, \phi)\text{a}_{00}, \tag{5}$$

and

$$\mathbf{y}_{\text{diff}} = \sqrt{\frac{\psi}{L}}\mathcal{D}[\mathbf{1}_L \text{a}_{00}], \tag{6}$$

[2] Assuming ortho-normalised (N3D) real SHs with ACN indexing. Omit the $1/\sqrt{3}$ term if using the semi-normalised (SN3D) convention.



(a) Components for sector 1

(b) Components for sector 2

(c) Components for sector 3

(d) Components for sector 4

Fig. 3. Example sector patterns using a 4-point t-design [43], second order spherical harmonic input, and hyper-cardioid beam-patterns. These are then treated as the pressure (first pattern on the left) and pressure gradient components (second, third and fourth patterns) to estimate four spatially localized active-intensity vectors (one per sector). The primary design philosophy of this approach is that parameters estimated in one sector should be less affected by noise and interferers present in the other sectors.

where $\mathbf{g}(\theta, \phi) \in \mathbb{R}^{L\times 1}$ are the VBAP gains corresponding to the estimated DoA; $\mathbf{1}_L \in \mathbb{R}^{L\times 1}$ is a vector of ones to replicate the omnidirectional component, a_{00}, to each loudspeaker channel; and $\mathcal{D}[.]$ denotes a de-correlation operation on the enclosed RIRs. The final time-domain loudspeaker array RIR may be acquired by summing the two streams, followed by the appropriate inverse time-frequency transform.

2.3 Higher-Order Spatial Impulse Response Rendering

The legacy SIRR formulation operates on only first-order spherical harmonic RIRs, but was still found to perform well for a variety of conditions in [4], despite its low input resolution. However, for problematic cases, such as those with multiple prominent early reflections arriving at the receiver position simultaneously from different directions, the method performs less optimally. This is primarily due to the limitations of the active-intensity based parameter estimation, as it is unable to distinguish between two reflections in the same time-frequency tile.

The new higher-order analysis conducted by the HO-SIRR method is based on first partitioning the sound-field into a number of uniformly distributed spatially localized sectors, S, which are then used to obtain spatially localized active-intensity vectors [40, 39]; examples of sector patterns are depicted in Fig. 3. As is the case with DoA estimates made using the traditional active-intensity vector Eq. (4), the DoA estimates extracted from sectors are also continuous, i.e., they forgo the need for dense scanning grids and peak-finding algorithms.

The spatial analysis is conducted for each sector individually in the same manner as before, yielding the following higher-order parameter vector

$$\mathbf{p}_N = \mathcal{A}\left[\mathbf{s} = \mathbf{W}_S \mathbf{a}_N\right] = [\theta_1, \phi_1, \psi_1, ..., \theta_S, \phi_S, \psi_S], \tag{7}$$

where $\mathbf{W}_S \in \mathbb{R}^{4S\times(N+1)^2}$ is a beamforming matrix which generates the spatially localized components for *all* sectors, $\mathbf{s} \in \mathbb{R}^{4S\times 1} = [\mathbf{z}^{(1)}; \mathbf{z}^{(2)}; ...; \mathbf{z}^{(S)}]$, where $\mathbf{z}^{(s)} = [z_{00}^{(s)}, z_{1(-1)}^{(s)}, z_{10}^{(s)}, z_{11}^{(s)}]$ are the components for each individual sector. For details regarding the formulation of $\mathbf{W}_S$, the reader is referred to [6, 40, 39].

The sector components are then panned to the loudspeaker channels using VBAP and accumulated, which yields the direct stream as

$$\mathbf{y}_{\text{dir}} = \sum_{s=1}^{S} \sqrt{\frac{1-\psi_s}{S}}\, \mathbf{g}(\theta_s, \phi_s) z_{00}^{(s)}. \tag{8}$$

For the diffuse stream, the sector signals are first scaled with the corresponding diffuseness values and re-encoded back into the SHD

$$\mathbf{a}_{\text{diff}} = \left\{ \begin{array}{ll} \sqrt{\frac{\psi_1}{(N+1)^2}}\mathbf{a}_1, & \text{for } N = 1 \\ \sum_{s=1}^{S} \sqrt{\frac{\psi_s}{S}}\mathbf{y}_{N-1}^{(s)} z_{00}^{(s)}, & \text{for } N > 1 \end{array} \right\}, \tag{9}$$

where $\mathbf{y}_{N-1}^{(s)}$ are the spherical harmonic weights for the corresponding sector direction, and $\mathbf{a}_{\text{diff}}$ is the diffuse stream expressed in the SHD.

The loudspeaker diffuse stream, $\mathbf{y}_{\text{diff}}$, is then obtained via linear Ambisonic decoding, followed by a decorrelation operation on the loudspeaker channels

$$\mathbf{y}_{\text{diff}} = \mathcal{D}\left[\mathbf{D}_{\text{ls}} \mathbf{a}_{\text{diff}}\right]. \tag{10}$$

Note that for $N > 1$, this diffuse stream rendering approach may reproduce the diffuse components in an anisotropic manner, since the sector components are scaled by direction dependent diffuseness values. Furthermore, if only first-order input is available then the method reduces

to the legacy SIRR analysis and synthesis, with the exception of employing an ambisonic decoder for the diffuse stream (rather than replicating the omnidirectional component to each loudspeaker channel). This is generally more preferable, as it allows for some degree of spatial separation between the loudspeaker channels, and thus, less aggressive decorrelation is required; as first suggested in [32].

3 LABORATORY TESTS FOR EVALUATING THE REPRODUCTION OF SPATIAL IMPULSE RESPONSES

The ultimate aim of any spatial RIR reproduction method is to reproduce a captured room response in such a manner, that the listener cannot distinguish it from the original. Therefore, when evaluating such methods, a listening subject should be able to directly compare the original room response with that of the reproduced response. However, as the human auditory memory is quite limited, direct comparison between the real space and its reproduction is generally not feasible. For development purposes, however, this may be circumvented by recreating reference scenarios with loudspeaker setups, which comprise similar components as experienced with natural scenarios, i.e., direct sound, reflections, and diffuse reverberation. Therefore, the purpose of this section is to describe the listening setup employed for this study and also the manner in which the reference scenarios were generated.

3.1 Listening Room Setup

In this study, the Audio-Visual Immersion Lab (AVIL) located at the Technical University of Denmark was utilized for reproducing the reference scenarios and auralizing the responses rendered by different methods. The audio system in AVIL comprises a spherical array of 64 loudspeakers (KEF, Maidstone, UK) placed in an anechoic chamber of approximate dimensions (length $\times$ width $\times$ height) 7 m $\times$ 8 m $\times$ 6 m. The listener is seated on a height-adjustable chair in the middle of the sphere at a distance of 2.4 m from the loudspeakers, which are arranged on seven concentric rings at elevation angles $\pm 80^\circ$, $\pm 56^\circ$, $\pm 28^\circ$, and 0°, with 2, 6, 12, and 24 loudspeakers, respectively, distributed on said rings. Three sonible d:24 amplifiers (sonible GmbH, Graz, Austria) drive the loudspeakers, and the digital-to-analogue conversion is performed by two biamp Tesira Server units (biamp Systems Inc., Beaverton, Oregon). Time, level, and magnitude response corrections were applied to each loudspeaker channel based on impulse response measurements made at the centre of the array.

3.2 Synthesis of Reference Loudspeaker Array RIRs

Acoustical simulation tools were used to create reference RIRs for each loudspeaker in the 64-channel spherical loudspeaker array. Essentially, the workflow involved first using modeling software to simulate different acoustical environments. Acoustic room modeling software, such as ODEON [45], CATT [46], and Ramsete [47], are capable of fulfilling this first task, as they may simulate the propagation of sound through virtual acoustic environments and also provide tools to manipulate important parameters (e.g., geometry and absorption coefficients), thus permitting the user to affect the overall acoustics of the modeled space.

In this study, ODEON (ODEON A/S, Lyngby, Denmark) was employed in combination with the Loudspeaker-based Room Auralization (LoRA) [48][3] toolbox. Echograms exported by ODEON, alongside the directional metadata, were passed to the LoRA toolbox, which employs dedicated rendering strategies for the early and late parts of the response. The early components are rendered directly to the loudspeaker RIRs based on this metadata, and the late parts are modeled as exponentially decaying noise sequences tuned to the directional reverberation times (RT60) in octave bands. Note that this workflow has previously been shown to be an ecologically valid approach for conducting listening tests in [49, 50].

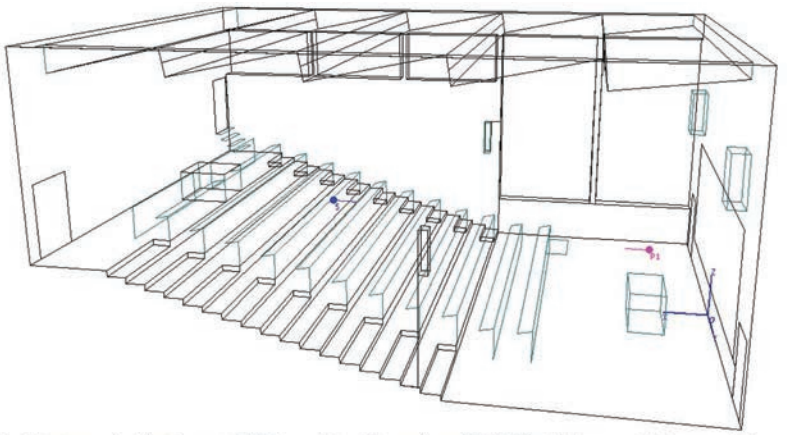
Once LoRA had rendered the reference 64-channel loudspeaker RIR for each room model, the spherical harmonic RIR counterparts were subsequently obtained via the application of a matrix of spherical harmonic encoding weights [1], with each column comprising the spherical harmonic coefficients for each loudspeaker direction in the AVIL array. This then allowed the spherical harmonic RIRs to be rendered to the same loudspeaker array using different rendering methods and configurations. Therefore, this workflow not only ensured that the rendering methods under test all shared the same input, but it also permitted direct comparisons between the different renders and the original loudspeaker RIRs generated by LoRA.

3.3 Room Models Used for Listening Tests

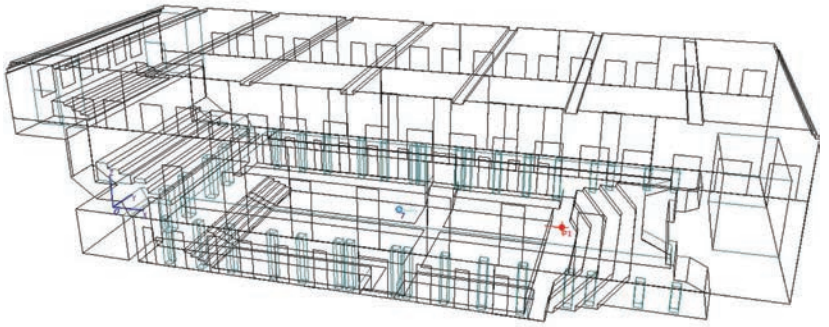
Two models of acoustic environments were used for the formal listening tests detailed in Sec. 5, namely the Vienna Musikverein concert hall and a lecture auditorium at DTU. Both were selected to be representative of the likely use cases for a spatial RIR reproduction method and exhibited different RT60s and room geometry. For instance, the Vienna Musikverein is a world-renowned venue for music performances, with dimensions 50.7 m $\times$ 19.5 m $\times$ 15 m, a volume of 13,337 m^3, and broad-band RT60 of 3.4 seconds. The source and receiver positions (x, y, z) were [36.79; 1.82; 2.16] m and [24.72; 0.4; 2.72] m, respectively, with a source-receiver distance of 12 m, whereas the DTU auditorium has dimensions 15.8 m $\times$ 12.1 m $\times$ 7.4 m, a volume of 1,177 m^3, and RT60 of 1.0 seconds. The source and receiver positions were [1.5; –1.5; 1.2] m and [10.3; –0.15; 2.55] m, respectively; with a source-receiver distance of 9 m. The geometries and source-receiver positions for these two acoustic environments are depicted in Fig. 4.

Additionally, during the informal listening described in Sec. 4, a room model of a small classroom at DTU was also used. Note that all of these models are available as example rooms in the ODEON software package. The "room

[3]The open-source LoRA toolbox may be found here: https://bitbucket.org/hea-dtu/lora/src/master/.



(a) Lecture Auditorium at DTU model. Dimensions (L,W,H): 15.8 m × 12.1 m × 7.4 m; Volume: 1,177 m^3; and RT60: 1.0 seconds.



(b) Vienna Musikverein concert hall model. Dimensions (L,W,H): 50.7 m × 19.5 m × 15 m; Volume: 13,337 m^3; and RT60: 3.4 seconds.

Fig. 4. The geometries of the ODEON models used for the simulations, including the source (magenta/red icon) and receiver (dark/light blue icon) positions. Note that both models are available as example rooms in the ODEON software package.

setup" parameters were defined using the *Precision* preset in ODEON, which was found to cast a sufficient number of rays for the authors' perceptual evaluation requirements. For each room, the source was defined as a "point" source and the receiver was omnidirectional.

4 INFORMAL LISTENING TESTS AND MULTICHANNEL ENERGY-SPECTROGRAM ILLUSTRATIONS

The subject matter was first approached through extensive informal observations in the AVIL listening space, where different rendering methods were compared against a reference scenario and with each other; using a variety of different acoustics and input stimuli. Simulations of an Auditorium, the Vienna Musikverein concert hall and a small classroom (as described in Sec. 3) were used as reference scenarios. Four different monophonic sound samples ("clicky" kick drum, trombone, speech, and castanets) were employed as the input stimuli. It should be noted that each of these sound samples exhibit unique spectral and temporal characteristics, and were specifically chosen to be revealing of any deficiencies present in the parametric rendering methods. These sound samples were subsequently convolved with the reference loudspeaker RIRs and also with the reproduced RIRs using SDM-B-Format [8] and various different configurations of the HO-SIRR method.

In the early phase of informal testing, it was found that short broad-band sound samples, such as the kick drum, were especially critical in revealing certain artifacts, such as reflections being perceived as too loud and/or the reverberant tail being found to be "grainy" (rather than exhibiting a smooth decaying response). On the other hand, sounds with ongoing tonal components and with narrower harmonic spectral content, largely masked these artifacts. Furthermore, the longer the reverberation time of the modeled room, the more ambiguous the differences between rendering methods became.

In the case of SDM-B-Format, particularly when using the Auditorium room with the kick drum sample, the reproductions greatly deviated from the reference scenario; often exhibiting considerable perceivable degradations. These findings appear to contradict the reproduction accuracy reported in [2]. However, there are potentially two key reasons

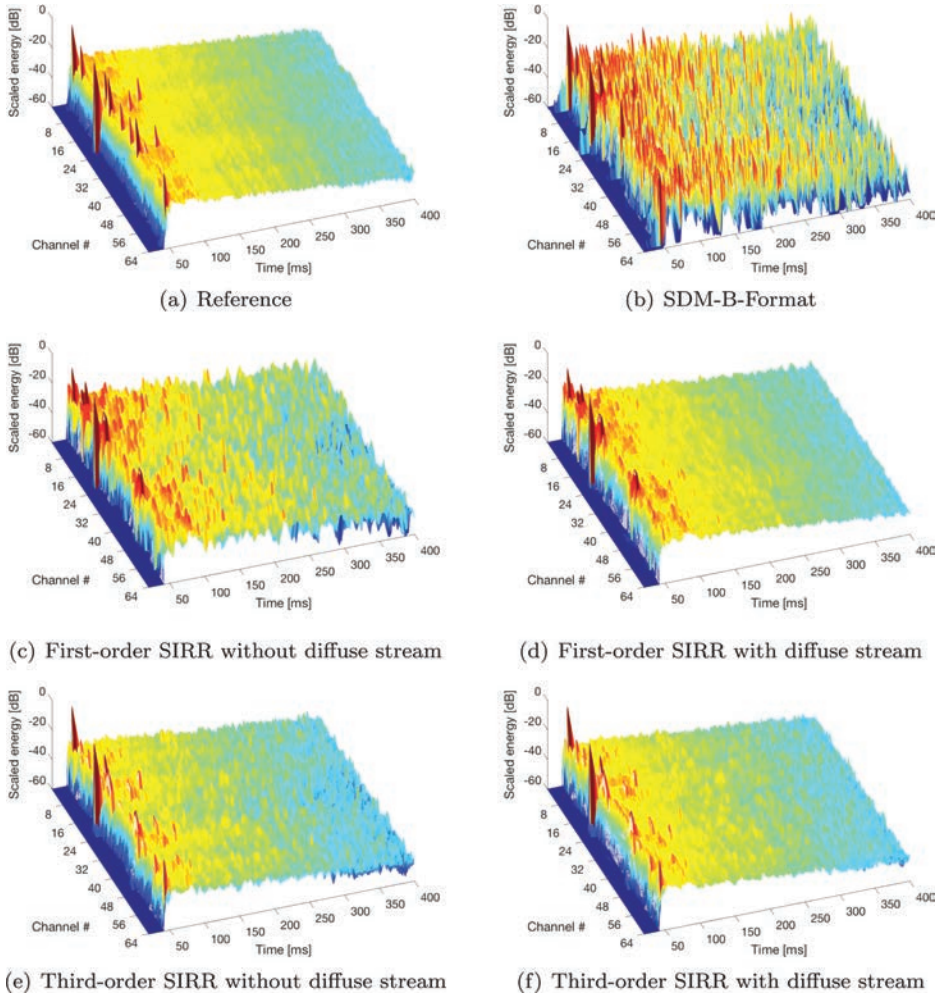


Fig. 5. Energy spectrograms of the Auditorium RIRs rendered for the 64-channel AVIL setup, using different methods and configurations; adopted from [37]. Note that the depicted SIRR cases employed the settings as detailed in Sec. 5. Additionally, the default configuration found in the SDM Matlab toolbox [8] was employed for the SDM-B-Format case, which essentially converges to a broad-band variant of SIRR without diffuse stream rendering.

for these differing outcomes. Firstly, these aberrations may have previously been masked due to the listening setup employed in [2], which comprised an acoustically treated (but still echoic) listening room and 14 loudspeakers in total. This was in contrast with the 64 loudspeakers and anechoic listening setup employed in this study. Therefore, the listening test subjects in [2], experienced the reproduced RIR in combination with the RIR of the listening room. Secondly, since SDM collapses the whole sound-field into a single loudspeaker direction for each time window, it is intuitively clear that (given a constant temporal window length): the more loudspeakers introduced into the array, the more SDM will have the tendency towards reproducing the late response too sparsely.

Regarding the HO-SIRR method, the present authors found that the inclusion of a dedicated diffuse stream renderer and higher-order processing, both noticeably improved the perceived reproduction accuracy for the challenging kick drum sample, whereas the inclusion of frequency-bands provided a less substantial (but still identifiable) improvement. However, the magnitude of these perceived improvements largely diminished for the more stationary trombone sample. Energy spectrograms of the reference loudspeaker RIR and the reproduced RIRs, as shown in Fig. 5, appear to support these informal findings. Note that the RIRs have been low-pass filtered with a first-order IIR filter with cut-off frequency of 100 Hz, prior to computing the energy spectrograms, in order to improve the graphical clarity. The plotting script and reference loudspeaker RIR are included in the HO-SIRR Matlab toolbox[4].

The reference response is shown in Fig. 5(a), where the first path and early reflections are clearly visible as distinct peaks, and the late reverberation has a relatively smooth downward-trending surface. The SDM-B-Format reproduction is depicted in Fig. 5(b), where the early peaks are shown to be reproduced correctly. However, the late part of the response comprises a multitude of peaks which are not found in the reference response. These additional peaks are likely the cause for the excess loudness of reflections and the graininess of the reverberant tail, as perceived by

[4]The open-source HO-SIRR Matlab toolbox may be found here: https://github.com/leomccormack/HO-SIRR.

Table 1. Test configurations. Test 1 studied the effect of frequency resolution and diffuse stream rendering, whereas Test 2 investigated the effect of spherical harmonic input order.

	Stimuli	Rooms	Conditions
Test 1	Kick drum Trombone	Auditorium	SDM B-Format o1 SIRR BB NoDiff o1 SIRR 6-ERB NoDiff o1 SIRR 128-bins NoDiff o1 SIRR BB o1 SIRR 6-ERB o1 SIRR 128-bins o1
Test 2	Kick drum Trombone Speech	Auditorium Vienna Musikverein	Ambisonics o1 Ambisonics o3 Ambisonics o5 SIRR 128-bins o1 HO-SIRR 128-bins o3 HO-SIRR 128-bins o5

the present authors. This would also appear to indicate that the primary SDM assumption (that the whole sound-field should be collapsed into a single direction at each time window) may become increasingly unrealistic as the density of reflections increases. Interestingly, when the analysis of the response is conducted in frequency bands without a dedicated diffuse stream, as depicted in Fig. 5(c), the response is slightly less sporadic. However, the late response is still more turbulent than it appears in the reference. The inclusion of a dedicated diffuse stream appears to render the surface of the late response in a smoother manner, as seen in Fig. 5(d). In the third-order cases, Fig. 5(e) and Fig. 5(f), the energy spectrograms become visibly more similar to the reference case, suggesting that the rendering performance is indeed improved with higher-order analysis and synthesis.

Note, however, that these figures and informal discussions are provided only to give an impression of the differences of various rendering methods and configurations. The perceptual consequences of these visible differences are investigated with formal listening tests in the following section.

5 FORMAL LISTENING TESTS

In order to further investigate the perceptual effects identified during informal listening of SDM-B-Format and the different rendering configurations of HO-SIRR (as described in Sec. 4), formal listening tests were conducted in the DTU AVIL listening room. A total of eight self-reported normal hearing male listeners participated, all of whom were naïve as to the hypothesis of the study. All listeners were experienced in spatial audio evaluation, and are employees within the Hearing Systems group at DTU.

5.1 Listening Test Interface

A multiple-stimulus listening test was designed, where the subject sat in the centre of the AVIL loudspeaker array. The HULTI-GEN MaxMSP project[5] by Gribben and Lee [51], was employed as the test user interface, and a touch-screen display was used to control the interface from the listening position. The listener was able to compare the reference to the reproductions and rate the similarity of their reproduction using a slider, which depicted the verbal anchors of the ITU scale: "bad", "poor", "fair", "good", and "excellent." The integer numerical values of the slider were 0–20, 21–40, 41–60, 61–80, and 81–100, respectively. Each listener was given an obligatory 10–15 minutes of training to familiarize themselves with the stimuli and to ensure they were comfortable interacting with the touch-display. Each trial was repeated once per stimulus (a total of two presentations per stimulus), and the presentation order of the trials was randomized. The listeners were encouraged to take short pauses when needed to avoid fatigue and were permitted to move their head while listening. The typical duration for the test (including training) was one hour.

5.2 Tested Effects

The listening test was divided into two parts. The first part investigated the effect of **frequency-resolution** and **dedicated diffuse stream rendering**. Here, the default configuration of the SDM-B-Format method, from the SDM Matlab toolbox [8], was employed as one of the test configurations, which uses first-order input (*o1*) and is labelled henceforth as *SDM-B-Format o1*. Note that the default SDM-B-Format configuration (at the time of writing) employed a temporal Hann window of length 0.3 ms (15 samples at 48 kHz) with 99% window overlap; quantized the pressure signal to the nearest loudspeaker; and had the adaptive post-equalization feature enabled (as described in [35]). The remaining test configurations were of first-order SIRR using the HO-SIRR Matlab toolbox, developed by the present authors. Note that (at the time of writing) the default settings comprised first-order analysis/synthesis with 50% overlapping temporal windows of length 5.$\dot{3}$ ms (256 samples at 48 kHz). In addition, VBAP [33] was employed as the panning method and the dedicated diffuse-stream rendering feature was enabled. The diffuse stream was decorrelated by convolving each channel with exponentially decaying independent Gaussian noise sequences, with the following decay rates

[5]The test interface is freely available from here: https://research.hud.ac.uk/institutes-centres/apl/resources/.

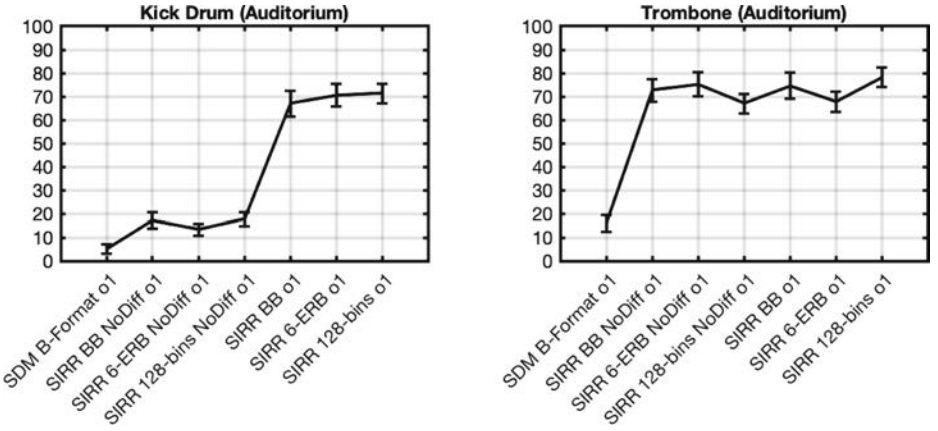


Fig. 6. The means and 95% confidence intervals of the results for the first listening test, which simultaneously investigated the perceptual differences of frequency-resolution and dedicated diffuse stream rendering.

per octave band (125 Hz to 4 kHz): [70, 70, 60, 40, 20, 10] ms. These decorrelation sequences were also equalized based on their minimum-phase representation, in order to make them spectrally flat, as described in [52]. The diffuseness parameter was computed based on an active-intensity vector averaged across frequency bins $\in$ (0, 3 kHz] and over time, using a one-pole averaging filter with a coefficient value of 0.975. The first peak in the input spatial RIR was also isolated and panned using a broad-band DoA estimate. Note that the toolbox employs the reformulated diffuse stream rendering approach, Eqs. (9) and (10), rather than the legacy diffuse rendering of Eq. (6).

This default HO-SIRR toolbox configuration was then kept constant for all of the SIRR cases for the first listening test, with two exceptions. Firstly, three different frequency resolutions were attained by appropriately averaging the intensity vector across the 128 (usable) frequency bins. This was done to keep the temporal resolution constant, while obtaining the following frequency resolution options: broad-band (*BB*), six Equivalent Rectangular Bandwidth (ERB) bands (*6-ERB*), and the full 128 uniformly spaced bins (*128-bins*). Secondly, the diffuse stream was disabled (*NoDiff*) for one combination of the three different frequency resolutions, and was enabled for the other combination of the three frequency resolutions. Therefore, there were seven test conditions in total (not including the reference). Furthermore, in this first test, stimuli "kick drum" and "trombone" were used in conjunction with the simulated "auditorium" room.

The third effect of interest in this study was the influence of **spherical harmonic input order** on the perceived accuracy of the reproduction. This was studied in part two of the listening test. Here, the HO-SIRR toolbox was configured with default settings (as before), except that the spherical harmonic input order was set to either first (*o1*), third (*o3*), or fifth (*o5*) order. Since HO-SIRR is currently the only parametric rendering method for spatial RIRs which supports higher-order input, the nonparametric *Ambisonics* method was included for comparison using the same input orders. Note that the mode-matching ambisonic decoder (MMD), as defined in Sec. 2.1, was utilized for this task. This decoder was also the ambisonic decoder used for the diffuse-stream rendering in the HO-SIRR toolbox. In total, this second test comprised six conditions. Furthermore, three stimuli ("kick drum," "trombone," and "speech") were used in conjunction with two room simulations: "auditorium" and "Vienna Musikverein." Note that the full listening test configurations are summarized in Table 1.

The subjects were then instructed, for both listening test parts, to rate the test cases based on how accurately they reproduced the reference case, particularly in regard to localization accuracy, envelopment, and timbre. However, it was also emphasized that other attributes of the reproduced stimuli may also be included in their judgement as they saw fit.

6 RESULTS AND DISCUSSION

The results for part 1 of the formal listening test, which investigated the perceived differences of **frequency resolution** and **dedicated diffuse rendering**, are presented in Fig. 6. When observing the results of using the kick drum sample in conjunction with the Auditorium model, it would appear that increasing the frequency-resolution yields little to no benefit. This is in contrast with the informal listening observations made by the present authors. However, the inclusion of dedicated diffuse-stream rendering indicates a substantial perceptual benefit. Therefore, it is possible that the perceived differences of changing the frequency resolution were overshadowed by the benefits of including the dedicated diffuse-stream rendering. When employing the trombone sample, which is a comparatively more stationary sound source, with the same room model, there appears to be little difference between both frequency resolution and dedicated diffuse-rendering. However, the SDM-B-Format test case, which should have some parity with the *SIRR BB NoDiff o1* case, was found to score much lower. There are four main differences between these two test cases: the temporal resolution (15 samples versus 256

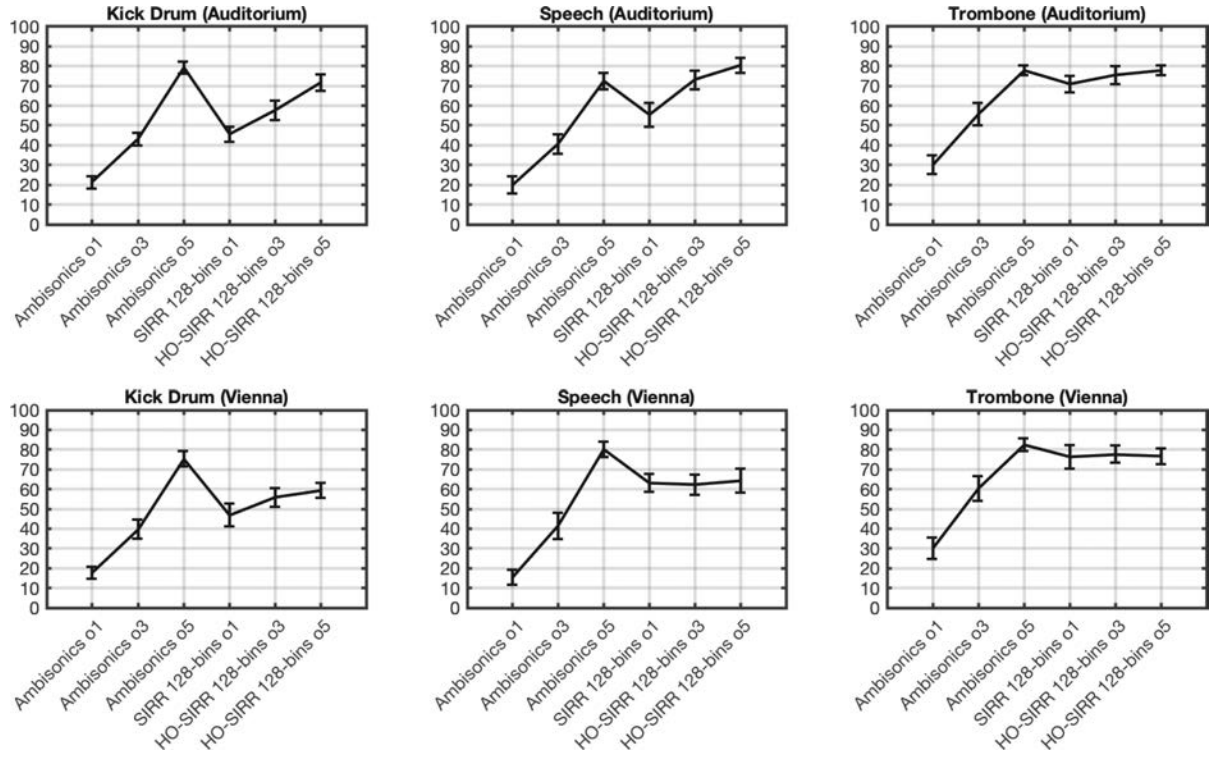


Fig. 7. The means and 95% confidence intervals of the results for the second listening test, which investigated the perceptual differences of spherical harmonic input order.

samples window length, at 48 kHz); window overlap (99% versus 50%); quantizing to the nearest loudspeaker rather than employing VBAP; and SDM-B-Format employs additional adaptive post-equalization [35]. It is not fully clear which of these attributes (or combination of these attributes) is responsible for the poor performance found in this study, although it is likely that the short temporal windowing and spatial quantization (instead of smooth panning functions) could contribute to a less smooth time-varying distribution of energy in the output. Furthermore, strong perceptual degradations compared to reference scenarios have previously been reported in [53], where the author cited the over simplification of the model and the heavy reliance on post-equalization as the root causes for the poor performance of SDM. However, it should also be noted that a recent binaural study [54] found that SDM formulated in the space domain (as originally intended [2]) can yield better performance than SDM formulated in the SHD.

The results for part two of the formal listening test, which investigated the perceptual effects of increasing the **spherical input order**, are depicted in Fig. 7. For the Auditorium room with the kick drum and speech stimuli, there appears to be a clear improvement in the perceived reproduction accuracy (for both methods) with increasing input order. When comparing the scores of Ambisonics alongside those of HO-SIRR, it would appear that the parametric processing substantially improves the perceived reproduction accuracy with first and third order input. With fifth order input, Ambisonics attained a slightly higher score for the kick drum sample, but fared worse with the speech sample; suggesting that there are some diminishing returns when employing HO-SIRR at very higher orders and with more transient input sources. However, since most commercially available microphone arrays are unable to capture these fifth-order components, the large improvements at first- and third-order (which represent more realistic input resolutions) are perhaps of greater practical significance. For the trombone case, the differences between input orders for the HO-SIRR method was minimal, as all three were largely indistinguishable from the fifth-order Ambisonics test case. Therefore, the performance improvement of higher-order processing appears to be largely negated when employing more stationary input samples. This also concurs with the informal observations detailed in Sec. 4.

When observing the results for the Vienna hall, the benefit of increasing the input order is less clear, but still present when employing the problematic kick drum sample. However, the benefits of employing parametric rendering, rather than linear Ambisonic decoding, appear to diminish above third-order input. One possible reason for this is that the default averaging coefficients and temporal window length in the HO-SIRR toolbox (at the time of writing), were selected through informal listening of the Auditorium room case. Therefore, it is possible that tailoring these parameters for the Vienna hall could yield results more in line with those of the Auditorium case. Therefore, further development of the HO-SIRR toolbox is a topic of future work. A future study could also comprise perceptual tests involving more rooms and test configurations, and investigating the binaural reproduction performance of the method. As a

final remark, while only eight subjects participated in the listening tests, it should be highlighted that each test configuration was repeated twice and that the confidence intervals are quite narrow, which suggests high reproducibility of the presented results.

7 CONCLUSIONS

This study investigated the perceptual effects of different design choices when parametrically rendering loudspeaker room impulse responses (RIRs), using spherical harmonic RIRs as input. More specifically, a listening setup comprising 64 loudspeakers in an anechoic chamber was employed to assess the perceptual effects of the following three aspects: frequency-resolution, dedicated diffuse stream rendering, and the input spherical harmonic order. The evaluation of these three different design choices was conducted through two formal listening tests.

The first listening test simultaneously investigated the perceptual effects of frequency-resolution and dedicated diffuse stream rendering. Here, renders using different configurations of the first-order Spatial Impulse Response Rendering (SIRR) method, and the spherical harmonic variant of the Spatial Decomposition Method (SDM), were compared against reference scenarios. The first-order SIRR configurations under test comprised three different frequency-resolutions, which had the dedicated diffuse stream rendering feature either enabled or disabled. The results indicate that by enabling this dedicated diffuse stream rendering feature, the perceived accuracy of the reproduction is substantially improved. On the other hand, increasing the frequency-resolution employed during rendering appears to yield little to no benefit. However, it is possible that the perceived differences between frequency-resolutions were overshadowed by the benefits of dedicated diffuse stream rendering. Indeed, through informal listening, the present authors were able to identify minor improvements with increasing frequency-resolution. Therefore, future work could entail a formal investigation into these perceptual differences via a dedicated listening test. Furthermore, the spherical harmonic variant of SDM was found to perform worse than all tested configurations of first-order SIRR.

A higher-order formulation of the SIRR method (HO-SIRR) was also introduced in this article. This new formulation is intended to improve upon the spatial accuracy of the original first-order method, by employing higher-order input to first segregate the sound-field into multiple sectors. The analysis and synthesis stages are then conducted for each sector individually, which allows it to more reliably reproduce multiple simultaneous reflections. The diffuse stream is also conducted in a manner that may more faithfully reproduce anisotropic diffuse-fields. Therefore, the second listening test investigated the perceptual effects of the third rendering configuration of interest: the spherical harmonic input order. For this test, first, third, and fifth-order renderings of HO-SIRR were compared alongside mode-matching Ambisonic decoding of the same input orders. The results indicate that (in the majority of cases), the reproduced responses more closely resembled the reference with increasing input order for both methods. However, more importantly, the benefits of the HO-SIRR parametric rendering is clearly demonstrated for both first and third order input and, to a lesser extent, also with fifth order input; especially when employing problematic transient stimuli.

8 ACKNOWLEDGMENT

This research has received funding from the Aalto University Doctoral School of Electrical Engineering and the Academy of Finland project no. 317341.

9 REFERENCES

[1] B. Rafaely. *Fundamentals of Spherical Array Processing*, vol. 8 (Springer, Berlin, Germany, 2015).

[2] S. Tervo, J. Pätynen, A. Kuusinen, and T. Lokki "Spatial Decomposition Method for Room Impulse Responses," *J. Audio Eng. Soc.*, vol. 61, no. 1/2, pp. 17–28 (2013 Mar.).

[3] J. Merimaa and V. Pulkki "Spatial Impulse Response Rendering I: Analysis and Synthesis,". *J. Audio Eng. Soc.*, vol. 53, no. 12, pp. 1115–1127 (2005 Dec.).

[4] V. Pulkki and J. Merimaa, "Spatial Impulse Response Rendering II: Reproduction of Diffuse Sound and Listening Tests," *J. Audio Eng. Soc.*, vol. 54, no. 1/2, pp. 3–20 (2006 Feb.).

[5] A. Wabnitz, N. Epain, A. McEwan, and C. Jin, "Upscaling Ambisonic Sound Scenes Using Compressed Sensing Techniques," *Proc. 2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 1–4 (2011 Oct.).

[6] A. Politis, J. Vilkamo, and V. Pulkki, "Sector-Based Parametric Sound Field Reproduction in the Spherical Harmonic Domain, *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 852–866 (2015 Aug.).

[7] A. Politis, S. Tervo, and V. Pulkki "Compass: Coding and Multidirectional Parameterization of Ambisonic Sound Scenes," *Proc. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6802–6806 (2018 Apr.).

[8] SDM Matlab Toolbox, https://se.mathworks.com/matlabcentral/fileexchange/56663-sdm-toolbox (accessed 2019-05-09).

[9] V. Valimaki, J. D. Parker, L. Savioja, J. O. Smith, and J. S. Abel "Fifty Years of Artificial Reverberation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 5, pp. 1421–1448 (2012 Jul.).

[10] J. Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization* (MIT Press, Cambridge, Massachusetts, 1997).

[11] O. Santala and V. Pulkki "Directional Perception of Distributed Sound Sources," *The Journal of the Acoustical Society of America*, vol. 129, no. 3, pp. 1522–1530 (2011 Mar.).

[12] D. Romblom, C. Guastavino, and P. Depalle "Perceptual Thresholds for Non-Ideal Diffuse Field Reverber-

ation," *The Journal of the Acoustical Society of America*, vol. 140, no. 5, pp. 3908–3916 (2016 Nov.).

[13] M. A. Gerzon "Periphony: With-Height Sound Reproduction," *J. Audio Eng. Soc.*, vol. 21 no. 1, pp. 2–10 (1973 Feb.).

[14] A. Solvang "Spectral Impairment of Two-Dimensional Higher Order Ambisonics," *J. Audio Eng. Soc.*, vol. 56, no. 4, pp. 267–279 (2008 Apr.).

[15] A. Avni, J. Ahrens, M. Geier, S. Spors, H. Wierstorf, and B. Rafaely "Spatial Perception of Sound Fields Recorded by Spherical Microphone Arrays With Varying Spatial Resolution," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 2711–2721 (2013 May).

[16] L. Yang and X. Bosun "Subjective Evaluation on the Timbre of Horizontal Ambisonics Reproduction," *Proc. 2014 International Conference on Audio, Language and Image Processing (ICALIP)*, pp. 11–15 (2014 Jul.).

[17] P. Stitt, S. Bertet, and M. van Walstijn "Off-Centre Localisation Performance of Ambisonics and HOA for Large and Small Loudspeaker Array Radii," *Acta Acustica united with Acustica*, vol. 100, no. 5, pp. 937–944 (2014 Oct.).

[18] O. Santala, H. Vertanen, J. Pekonen, J. Oksanen, and V. Pulkki, "Effect of Listening Room on Audio Quality in Ambisonics Reproduction," presented at the *126th Convention of the Audio Engineering Society* (2009 May), convention paper 7664.

[19] S. Braun and M. Frank "Localization of 3D Ambisonic Recordings and Ambisonic Virtual Sources," presented at the *1st International Conference on Spatial Audio,(Detmold)* (2011 Nov.).

[20] S. Bertet, J. Daniel, E. Parizet, and O. Warusfel "Investigation on Localisation Accuracy for First and Higher Order Ambisonics Reproduced Sound Sources," *Acta Acustica united with Acustica*, vol. 99, no. 4, pp. 642–657 (2013 Jul.).

[21] F. Menzer, C. Faller, and H. Lissek "Obtaining Binaural Room Impulse Responses From B-Format Impulse Responses Using Frequency-Dependent Coherence Matching," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 396–405 (2011 Feb.).

[22] M. Frank and F. Zotter, "Spatial Impression and Directional Resolution in the Reproduction of Reverberation," *Fortschritte der Akustik, DAGA* (2016 Mar.).

[23] D. Romblom, P. Depalle, C. Guastavino, and R. King "Diffuse Field Modeling Using Physically-Inspired Decorrelation Filters and B-Format Microphones: Part I Algorithm," *J. Audio Eng. Soc.*, vol. 64, no. 4, pp. 177–193 (2016 Apr.).

[24] P. Coleman, A. Franck, P. Jackson, R. Hughes, L. Remaggi, and F. Melchior "Object-Based Reverberation for Spatial Audio," *J. Audio Eng. Soc.*, vol. 65, no. 1/2, pp. 66–77 (2017 Jan.).

[25] M. Zaunschirm, M. Frank, and F. Zotter, "BRIR Synthesis Using First-Order Microphone Arrays," presented at the *144th Convention of the Audio Engineering Society* (2018 May), convention paper 9944.

[26] J. Ahrens "Auralization of Omnidirectional Room Impulse Responses Based on the Spatial Decomposition Method and Synthetic Spatial Data," *Proc. 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 146–150 (2019 May).

[27] Y. Yamasaki and T. Itow "Measurement of Spatial Information in Sound Fields by Closely Located Four Point Microphone Method," *Journal of the Acoustical Society of Japan (E)*, vol. 10, no. 2, pp. 101–110 (1989 Jan.).

[28] G. Schiffrer and D. Stanzial "Energetic Properties of Acoustic Fields," *The Journal of the Acoustical Society of America*, vol. 96, no. 6, pp. 3645–3653 (1994 Jun.).

[29] D. Stanzial, N. Prodi, and G. Schiffrer "Reactive Acoustic Intensity for General Fields and Energy Polarization," *The Journal of the Acoustical Society of America*, vol. 99, no. 4, pp. 1868–1876 (1996).

[30] M. Karjalainen, J. Merimaa, T. Peltonen, and T. Lokki, "Measurement, Analysis, and Visualization of Directional Room Responses," presented at the *111th Convention of the Audio Engineering Society* (2001 Sep.), convention paper 5449.

[31] F. J. Fahy and V. Salmon "Sound Intensity," *The Journal of the Acoustical Society of America*, vol. 88, no. 4, pp. 2044–2045 (1990).

[32] V. Pulkki "Spatial Sound Reproduction With Directional Audio Coding," *J. Audio Eng. Soc.*, vol. 55, no. 6, pp. 503–516 (2007 Jun.).

[33] V. Pulkki "Virtual Sound Source Positioning Using Vector Base Amplitude Panning," *J. Audio Eng. Soc.*, vol. 45, no. 6, pp. 456–466 (1997 Jun.).

[34] J. Pätynen, S. Tervo, and T. Lokki "Amplitude Panning Decreases Spectral Brightness With Concert Hall Auralizations," presented at the *AES 55th International Conference: Spatial Audio* (2014 Aug.), conference paper P-13.

[35] S. Tervo, J. Pätynen, N. Kaplanis, M. Lydolf, S. Bech, and T. Lokki "Spatial Analysis and Synthesis of Car Audio System and Car Cabin Acoustics With a Compact Microphone Array," *J. Audio Eng. Soc.*, vol. 63, no. 11, pp. 914–925 (2015 Nov.).

[36] S. Moreau, J. Daniel, and S. Bertet, "3D Sound Field Recording With Higher Order Ambisonics–Objective Measurements and Validation of Spherical Microphone," presented at the *120th Convention of the Audio Engineering Society* (2006 May), convention paper 6857.

[37] L. McCormack, A. Politis, O. Scheuregger, and V. Pulkki "Higher-Order Processing of Spatial Impulse Responses," presented at the *23rd International Congress on Acoustics (ICA)* (2019).

[38] A. Politis, L. McCormack, and V. Pulkki, "Enhancement of Ambisonic Binaural Reproduction Using Directional Audio Coding With Optimal Adaptive Mixing," presented at the *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (2017).

[39] L. McCormack, S. Delikaris-Manias, A. Politis, D. Pavlidi, A. Farina, D. Pinardi, and V. Pulkki "Applications

of Spatially Localized Active-Intensity Vectors for Sound-Field Visualization," *J. Audio Eng. Soc.*, vol. 67, no. 11, pp. 840–854 (2019 Nov.).

[40] A. Politis and V. Pulkki "Acoustic Intensity, Energy-Density and Diffuseness Estimation in a Directionally-Constrained Region," *arXiv:1609.03409* (2016 Sep.).

[41] F. Zotter, H. Pomberger, and M. Noisternig "Energy-Preserving Ambisonic Decoding," *Acta Acustica united with Acustica*, vol. 98, no. 1, pp. 37–47 (2012 Jan.).

[42] F. Zotter and M. Frank "All-Round Ambisonic Panning and Decoding," *J. Audio Eng. Soc.*, vol. 60, no. 10, pp. 807–820 (2012 Oct.).

[43] R. H. Hardin and N. J. A. Sloane "McLaren's Improved Snub Cube and Other New Spherical Designs in Three Dimensions," *Discrete & Computational Geometry*, vol. 15, no. 4, pp. 429–441 (1996).

[44] H. Hacıhabiboğlu "Theoretical Analysis of Open Spherical Microphone Arrays for Acoustic Intensity Measurements," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 465–476 (2014 Feb.).

[45] C. L. Christensen *Odeon Version 15.13 Room Acoustics Program: Version 15.13, Industrial, Auditorium and Combined Editions* (Technical University of Denmark, Kongens Lyngby, Denmark, 2003).

[46] B.-I. L. Dalenbäck "Room Acoustic Prediction Based on a Unified Treatment of Diffuse and Specular Reflection," *The Journal of the Acoustical Society of America*, vol. 100, no. 2, pp. 899–909 (1996).

[47] A. Farina and L. Tronchin "3D Sound Characterisation in Theatres Employing Microphone Arrays," *Acta acustica united with Acustica*, vol. 99, no. 1, pp. 118–125 (2013 Jan.).

[48] S. E. Favrot, J. Buchholz, and T. Dau *A Loudspeaker-Based Room Auralization System for Auditory Research* (Technical University of Denmark, Kongens Lyngby, Denmark, 2010).

[49] S. Favrot and J. M. Buchholz, "Validation of a Loudspeaker-Based Room Auralization System Using Speech Intelligibility Measures," presented at the *126th Convention of the Audio Engineering Society* (2009 May), convention paper 7763.

[50] J. Cubick and T. Dau "Validation of a Virtual Sound Environment System for Testing Hearing Aids," *Acta Acustica united with Acustica*, vol. 102, no. 3, pp. 547–557 (2016 May.).

[51] C. Gribben and H. Lee, "Towards the Development of a Universal Listening Test Interface Generator in Max," presented at the *138th Convention of the Audio Engineering Society* (2015 May), convention paper .

[52] M. J. Hawksford and N. Harris "Diffuse Signal Processing and Acoustic Source Characterization for Applications in Synthetic Loudspeaker Arrays," presented at the *112th Convention of the Audio Engineering Society* (2002 Apr.), convention paper 5612.

[53] C. Hold *Spatial Decomposition Method on Non-Uniform Reproduction Layouts,* Master's thesis, Institut für Kommunikation und Sprache Fachgebiet Audiotechnologie, Technische Universität Berlin (2019).

[54] J. Ahrens "Perceptual Evaluation of Binaural Auralization of Data Obtained From the Spatial Decomposition Method," *Proc. 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 65–69 (2019 Oct.).

THE AUTHORS


Leo McCormack


Ville Pulkki


Archontis Politis


Oliver Scheuregger


Márton Marschall

Leo McCormack received his M.Sc. degree in Computer Communications and Information Sciences, majoring in Acoustics and Audio Technology, at Aalto University, Helsinki, Finland, and his B.Sc. in Music Technology and Audio Systems at the University of Huddersfield, Huddersfield, UK. He was a student intern at Fraunhofer IIS, Erlangen, Germany, in 2013–2014 and received an AES Convention best peer-reviewed paper award in 2018. He is currently a doctoral candidate for the Department of Signal Processing and Acoustics at Aalto University, Finland, researching parametric spatial audio technologies. His research interests include signal processing for sound-field manipulation and reproduction, microphone and hydrophone array processing, and sound source localization. Leo McCormack is also a lead developer for several open-source software projects related to spatial audio and is a member of the Audio Engineering Society and IEEE Signal Processing Society.

•

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.

•

Archontis Politis obtained his M.Eng. degree in Civil Engineering from Aristotle University, Thessaloniki, Greece, and his M.Sc. degree in Sound & Vibration Studies from the Institute of Sound and Vibration Research (ISVR), Southampton, UK, in 2006 and 2008, respectively. From 2008 to 2010, he worked as a graduate acoustic consultant in Arup Acoustics, UK, and as a researcher in a joint collaboration between Arup Acoustics and the Glasgow School of Arts on architectural auralization using spatial sound techniques. In 2016, he obtained a Doctor of Science degree on parametric spatial sound recording and reproduction from Aalto University, Helsinki, Finland. He also completed an internship at the Audio and Acoustics Research Group of Microsoft Research during the summer of 2015. He received an AES Convention best student paper award in 2013. He has held workshops on parametric spatial audio processing, and he has co-authored a book on the topic, published in 2017. He is currently a postdoctoral researcher at Tampere University, Tampere, Finland. His research interests include spatial audio technologies, virtual acoustics, array signal processing, and acoustic scene analysis.

•

Oliver Scheuregger was born in 1991 in Maidstone, UK. He graduated from the Tonmiester (Music and Sound Recording) degree program at Surrey University in 2013 and then started a full-time position at Fraunhofer IIS, Erlangen, where he continues to work to this day as an audio engineer in the binaural rendering group.

In July 2019, he also received his M.Sc in Engineering Acoustics from the Danish Technical University, Kongens Lyngby, Denmark. His thesis focused on sound texture perception in hearing-impaired listeners, a topic he is still researching today at DTU. He enjoys riding and fixing bikes.

•

Márton Marschall is a native of Budapest, Hungary. He received an M.Sc. degree in Electrical Engineering from Budapest University of Technology and Economics, Budapest, Hungary, in 2006, and an M.Sc. degree in Engineering Acoustics and a Ph.D. in electrical engineering from the Technical University of Denmark, Kongens Lyngby, Denmark, in 2008 and 2014, respectively.

Since 2014, he has been a postdoctoral and then senior researcher with the Centre for Applied Hearing Research, Technical University of Denmark. His research interests include spatial audio recording and reproduction, virtual environments, spatial hearing, and auditory modeling.