Spatial Analysis and Synthesis Methods: Subjective and Objective Evaluations Using Various Microphone Arrays in the Auralization of a Critical Listening Room

Alan Pawlak, Hyunkook Lee, Aki Mäkivirta, and Thomas Lund

***Abstract*—Parametric sound field reproduction methods, such as the Spatial Decomposition Method (SDM) and Higher-Order Spatial Impulse Response Rendering (HO-SIRR), are widely used for the analysis and auralization of sound fields. This paper studies the performance of various sound field reproduction methods in the context of the auralization of a critical listening room, focusing on fixed head orientations. The influence on the perceived spatial and timbral fidelity of the following factors is considered: the rendering framework, direction of arrival (DOA) estimation method, microphone array structure, and use of a dedicated center reference microphone with SDM. Listening tests compare the synthesized sound fields to a reference binaural rendering condition, all for static head positions. Several acoustic parameters are measured to gain insights into objective differences between methods. All systems were distinguishable from the reference in perceptual tests. A high-quality pressure microphone improves the SDM framework's timbral fidelity, and spatial fidelity in certain scenarios. Additionally, SDM and HO-SIRR show similarities in spatial fidelity. Performance variation between SDM configurations is influenced by the DOA estimation method and microphone array construction. The binaural SDM (BSDM) presentations display temporal artifacts impacting sound quality.**

***Index Terms*—Spatial audio, binaural rendering, spatial decomposition method (SDM), higher-order spatial impulse response rendering (HO-SIRR), binaural room impulse responses (BRIR), auralization, microphone arrays, subjective audio evaluation, mushra, direction of arrival (DOA), time difference of arrival (TDOA), pseudo intensity vectors (PIV).**

Received 12 December 2023; revised 28 May 2024 and 22 July 2024; accepted 12 August 2024. Date of publication 23 August 2024; date of current version 12 September 2024. This work was supported in part by Genelec Oy and in part by the University of Huddersfield. The associate editor coordinating the review of this article and approving it for publication was Dr. Jens Ahrens. *(Corresponding author: Alan Pawlak.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by University of Huddersfield, School of Computing and Engineering School Research Ethics and Integrity Committee under Application No. 1561875, and performed in line with the Project Ethical Review Form.

Alan Pawlak and Hyunkook Lee are with the Applied Psychoacoustics Lab (APL), University of Huddersfield, HD1 3DH Huddersfield, U.K. (e-mail: alan.pawlak@hud.ac.uk; h.lee@hud.ac.uk).

Aki Mäkivirta and Thomas Lund are with the Genelec OY, 74100 Iisalmi, Finland (e-mail: aki.makivirta@genelec.com; thomas.lund@genelec.com).

The material includes additional information on the Eigenmike em32 microphone array, spectral analyses of HRIRs and pressure signals used in this study, as well as detailed configurations for test systems with MATLAB code snippets. Contact alan.pawlak@hud.ac.uk for further questions about this work.

This article has supplementary downloadable material available at https://doi.org/10.1109/TASLP.2024.3449037, provided by the authors.

Digital Object Identifier 10.1109/TASLP.2024.3449037

I. Introduction

CRITICAL listening rooms provide an acoustically controlled environment for audio production and evaluation, adhering to recommendations that specify criteria such as reverberation time and operational room response curve [1]. Advances in audio processing enable auralization [2], virtually reproducing acoustic environments such as critical listening rooms, which may be beneficial for spatial audio reproduction setups such as Dolby Atmos [3] or ITU-R BS.2051-3 [4]. This eliminates the need for a physical room with complex multi-loudspeaker configurations, increasing accessibility to optimized listening conditions. However, current virtual acoustic methods have limitations in accurately modeling early reflections and acoustic parameters, producing plausible but inauthentic auralizations compared to measured references [5].

Alternatively, auralization can be achieved through spatial reproduction methods: (i) non-parametric (e.g., Ambisonics), assuming a linear, time-invariant mapping between the microphone array and the target reproduction system, requiring only array and playback system specifications; or (ii) parametric, employing signal-dependent spatial analysis prior to the processing. Parametric methods extract metadata such as direction of arrival (DOA) and magnitude of individual sound events from actual measurements and use it to synthesize the sound field. Unlike virtual acoustic methods that simulate metadata, parametric methods rely on real-world data, making them a potentially more faithful methodology for research into the perceptual impact of room reflections, as demonstrated by [6], [7]. Such research may enable the development of optimized auralization systems.

Over the past two decades, the principles of spatial analysis have driven significant growth in parametric spatial audio rendering. This trend began with the introduction of the Spatial Impulse Response Rendering (SIRR) [8], which utilizes first-order spherical harmonics (SHs) for sound field analysis and synthesis in the time-frequency domain. A subsequent, more straightforward approach known as the Spatial Decomposition Method (SDM) was introduced by Tervo et al. [9]. This method operates in the time domain and interprets each sample in

an impulse response as an image source characterized by both pressure and direction. The public availability of SDM as a MATLAB toolbox [10] has made it a popular choice for analyzing enclosed spaces, auralization, and research [6], [7], [11], [12], [13], [14], [15]. As object-based audio gained prominence, the Reverberant Spatial Audio Object (RSAO) was developed, parameterizing spatial room impulse response (SRIR) into a concise set of coefficients. This was aimed at enabling reverberation synthesis within audio object renderers [16].

Later, HO-SIRR, a higher-order adaptation of the SIRR, was introduced [17], offering enhanced spatial resolution through the use of higher-order SHs. Concurrently, the Ambisonic SDM (ASDM) was introduced, enabling the upscaling of First-Order Ambisonics (FOA) to Higher-Order Ambisonics (HOA) [18]. This approach offers several advantages. Tetrahedral arrays enable efficient capture of SRIRs, which can be easily encoded into the FOA. Subsequently, ASDM allows for upscaling the FOA-encoded SRIRs to the desired Ambisonic order, providing flexible control over the directional resolution [18], [19]. Ambisonic format also facilitates the integration of head-tracking and complements existing HOA workflows by enabling the convolution of upscaled HOA RIRs with audio stimuli for use in HOA mixes. Subsequent innovations include the binaural versions of SDM (BSDM) [20] and HO-SIRR [21], Four-Directional ASDM (4D-ASDM) [22], and the Reproduction and Parameterization of Array Impulse Responses (REPAIR) [23].

Despite the rapid developments, the SDM and SIRR (and HO-SIRR) remain predominant. Previous studies [9], [17], [23], [24] evaluated these methods primarily through subjective experiments with loudspeakers in virtualized environments. Simulations such as the Image Source Method (ISM) [25] eliminate microphone array imperfections and provide direct references for loudspeaker reproduction [9], [17], [24], but may not fully represent real-world conditions and inherently favor techniques such as SDM due to shared assumptions. Additionally, ideal SHs can be obtained in simulations, which is challenging with real spherical microphone arrays (SMAs), as shown in the supplementary material, Section S.I. Headphone-based designs enable evaluations under real-world conditions using measured binaural [18], [26] or loudspeaker [20] references. This approach provides controlled, reproducible conditions, better representing practical performance. However, it requires head tracking implementation and dense binaural room impulse response (BRIR) measurements to account for natural head movements, providing dynamic binaural cues. In contrast, loudspeaker reproduction inherently accounts for head movements without additional considerations.

Inconsistencies arise, such as those between the SDM and HO-SIRR evaluations, which may stem from different microphone array configurations or DOA estimation methods used. While SDM has been used with various microphone arrays, only two studies have examined how the array affects SDM's auralization quality [26] and DOA accuracy [20]. Despite commonly using signals from any omnidirectional microphone in arrays, no perceptual differences were found between signals from Ambisonics and a central omni microphone, although the research lacked methodological detail [27]. Furthermore, previous studies focused on evaluations in more reverberant spaces, with only one involving a production studio [18], and employed a limited subset of the systems of interest here. No comprehensive comparison of all these methods under real-world conditions has been done previously, which this study intends to do.

In this paper[1], we present a comprehensive perceptual evaluation to test the hypothesis that different synthetic BRIRs produced using various parametric spatial audio rendering techniques—specifically SDM, BSDM, and HO-SIRR—will differ in spatial and timbral fidelities compared to a reference BRIR recorded with the KU100 dummy head. Building on [9] and [26], we investigate whether a greater number of sensors in an array leads to improved localization performance for the SDM variant with TDOA-based DOA estimation. Notably, higher sensor counts can also benefit methods such as HO-SIRR, which leverage higher-order SHs to localize multiple simultaneous reflections and better model the diffuse sound field component [17]. To this end, we employ both an Eigenmike em32—equipped with 32 omni capsules on a rigid sphere with a 42mm radius—and a compact microphone array with six omni microphones. The Eigenmike em32 is used for HO-SIRR, while both arrays are used for SDM. Further, we examine the impact of DOA estimation algorithms in SDM, focusing on those based on time difference of arrival (TDOA) and pseudo-intensity vectors (PIVs), and assess the influence of using a dedicated omnidirectional microphone at the array's center as a pressure signal in SDM.

The study is confined to a room that complies with recommendations in ITU-R BS.1116-3 [1] since our main focus is on the spatial analysis and auralization of critical listening or sound mixing room. We operate under the assumption that our findings will be replicable in similar rooms adhering to this widely recognized standard. Our study utilizes source positions from six orientations, aligned with industry standards such as the ITU-R BS.2051-2 [4] and Dolby Atmos 7.1.4 configurations. The aim is to ground the discussion in scientific rigor amidst the rapid expansion of parametric spatial audio reproduction techniques, delineating current standings, identifying gaps, and defining a system for optimal spatial data capture and accurate reproduction in the context of critical listening room.

The paper is structured as follows: Section II introduces the DOA estimation methods, with a focus on TDOA and sound intensity vectors (SIVs), which are crucial to the SDM and SIRR methods. Section III elaborates on the specifics of the SDM and SIRR methods. Section IV details our methodology and experimental design. Section V showcases our perceptual study results, based on the MUSHRA methodology. Section VI reports on objective metrics. Section VII synthesizes our key findings. Section VIII concludes the paper.

II. Direction of Arrival Estimation Methods

Estimating the DOA is a fundamental aspect of parametric sound field reproduction methods, as it determines the wave's

[1]This paper extends our initial study [28], offering a detailed analysis with more subjects and source positions, objective metrics, in-lab experiments (replacing the previous remote setup due to COVID-19), refined Eigenmike em32 impulse response measurements, and revising evaluated systems.

origin and propagation direction. Various methods have been developed for DOA estimation, with comprehensive tutorials available in [29]. These methods include TDOA-based methods such as the Generalized Cross Correlation (GCC), subspace methods such as Estimation of Signal Parameters via Rotational Invariance Techniques (ESPRIT) or Multiple Signal Classification (MUSIC), beamforming approaches, and sound intensity vector (SIV)-based methods. Among these, the TDOA, SIV, and ESPRIT estimators are notably more computationally efficient than MUSIC and beamforming methods, as they do not rely on a scanning grid [29], [30]. This section will focus on two approaches particularly relevant to SDM and SIRR: TDOA and SIVs.

A. Time Difference of Arrival (TDOA)

The TDOA is frequently used to determine the source's DOA. It measures the time lag of a signal across multiple sensors. Knowing the sensors' relative positions and these time lags allows estimation of the source's origin direction. Common TDOA estimation techniques include cross-correlation with weightings such as GCC-SCOT and GCC-PHAT [31].

While the SDM can be combined with any appropriate DOA estimation algorithm, the foundational paper utilized the least squares method for DOA estimation via TDOA using GCC with no weighting [9]. As this algorithm was incorporated into a popular toolbox [10], it is often linked with the original SDM. For detailed equations and further technical specifics, please refer to [9], [31], [32], [33].

B. Sound Intensity Vector (SIV)

SIV-based DOA estimation offers a viable alternative to TDOA-based methods. The use of SIV for this purpose dates back more than two decades [34], [35]. SIV is the product of pressure and particle velocity, with its real and imaginary parts representing active (sound energy flow in the propagation direction) and reactive intensity (non-propagating sound energy), respectively [36]. SIRR estimates the DOA from the active intensity vector (AIV) derived from the zeroth and first-order eigenbeams in the time-frequency domain [8]. Jarrett et al. [37] introduced the term Pseudo Intensity Vector (PIV), synonymous with the AIV in the time domain, which has become widely adopted in the field.

The results of subjective studies by McCormack et al. [17] and Ahrens [38] suggested that broadband DOA estimation using PIVs is inferior to other SDM and SIRR configurations in auralization. However, Zaunschirm et al. demonstrated that using this method for frequencies between 200Hz and just below the microphone array's spatial aliasing frequency resulted in SDM rendering nearly identical to the binaural reference in terms of auditory image width, distance, and diffuseness [18]. Bassuet [39] emphasized minimizing microphone directivity effects in broadband PIV-based DOA estimation, proposing a 100Hz–5kHz filter range for consistent directivity characteristics in the Soundfield first-order Ambisonics microphone. For the Eigenmike em32, this range would lie approximately between 140Hz and 8kHz.[2]

[2]We compare the spatial correlation and level difference between ideal and Eigenmike em32-derived eigenbeams to identify the operational frequency range for this microphone array in the supplementary material, Section S.I.

Two recent methods, Spatially Localized Active Intensity Vectors (SL-AIVs) [40] and the dual-directional estimator (DDE) [41], have been developed to handle multiple sound sources using SIVs. SL-AIVs are employed in HO-SIRR. They utilize higher-order SHs and beamforming to partition the sound field, enabling DOA estimation for multiple simultaneous sources. DDE, on the other hand, can estimate two simultaneous directions using only the first-order SH.

III. Parametric Sound Reproduction Methods

A. Spatial Decomposition Method (SDM)

The SDM, introduced in 2013, utilizes the image source model for parametric sound field reproduction, treating impulse response samples as broadband image sources [9]. The resulting metadata can be utilized for loudspeaker reproduction or binaural reproduction using Head-Related Transfer Functions (HRTFs). SDM's process involves a microphone array for DOA estimation and an omnidirectional microphone for pressure signal capture. The method comprises two stages: spatial analysis using DOA estimation algorithms (TDOA or PIVs) and synthesis, which utilizes DOA data and pressure signals to create directional output signals using techniques such as Vector Base Amplitude Panning (VBAP) [9], [14], [42], K-Nearest Neighbour (KNN) mapping [12], [38], [42], or Ambisonics [18], [19]. For scenarios prioritizing rendering quality over dynamic listener interaction, direct binaural rendering using densely sampled HRTFs provides high fidelity reproduction [18]. Regardless of the rendering approach, post-equalization to mitigate spectral whitening is commonly applied [9], [12], [18], [19]. Optimized virtual loudspeaker grids are utilized for enhancing binaural reproduction quality when KNN-based rendering is used [43].

Binaural Spatial Decomposition Method (BSDM) offers improvements for binaural reproduction [20]. It includes the rotation of the DOA matrix for various head orientations and post-processing techniques for direct sound enforcement and DOA quantization. BSDM also includes a reverberation correction process (RTMod) and a post-processing method that applies a cascade of allpass filters to the synthesized BRIRs (RTMod+AP). These techniques improve echo density and decay, building upon the previous post-equalization method designed for loudspeaker reproduction [12]. Later, updates to the publicly available toolbox have introduced features such as impulse response denoising and band-limited spatial analysis [44].

SDM's performance hinges on the microphone array configuration and window size. Historically, SDM has employed a variety of microphone arrays, from GRAS 50VI probes to custom arrays with varying spacings [9], [11], [12], [13], [20], [26], [27], [45], [46]. Optimal array design, guided by research, focuses on specific microphone spacings to minimize DOA errors and perceptual discrepancies [9], [20], [26]. PIV-based DOA estimation allows flexibility in microphone array choice, but it requires an array capable of producing high-quality first-order SHs [20]. In this method, windowing aims at smoothing the DOA, while in TDOA-based estimation, the window size is key to performance, ensuring a balance between temporal and spatial resolution. Larger windows enhance estimation robustness but increase the risk of multiple reflections arriving within the same

time window, conflicting with the single-reflection assumption of the SDM [9]. Historically, window sizes were chosen arbitrarily [11], [12], [42], [45]. Recent research by Amengual Garí et al. [20] on optimal window size for DOA estimation in a simulated setup, suggests that for a 100mm spaced array, a window length of 36 or 64 samples is most effective at a sampling rate of 48kHz.

B. Spatial Impulse Response Rendering (SIRR)

The SIRR method was the first parametric approach proposed for SRIR reproduction [8], [24]. Distinct from the SDM, SIRR operates in the time-frequency domain using first-order SH input and assumes a sound field model consisting of a single time-varying directional sound event per frequency band and an isotropic diffuse field.

The processing in SIRR utilizes the short-time Fourier transform with a Hann window, using AIVs—as outlined in Section II-B—to determine the DOA and diffuseness coefficient for each time-frequency bin. The diffuseness is estimated as the ratio of the magnitude of the AIV to the total sound energy within an analysis window [47]. As more reflections start to land in the same time window, the diffuseness coefficient tends towards 1. This shift moves the method away from single-source rendering and towards rendering everything with its diffuse component renderer. This is the key distinction of SIRR compared to SDM, which does not consider explicit diffuse field rendering and collapses the entire sound energy to a single point at a given time.

The synthesis stage divides the omnidirectional pressure signal into non-diffuse and diffuse streams based on the estimated diffuseness. The directional part is panned using VBAP, while the diffuse component is decorrelated and distributed uniformly among the loudspeakers.

Higher-Order Spatial Impulse Response Rendering (HO-SIRR) [17] extends the SIRR method to higher-order SH input, enabling the sound field to be divided into uniformly distributed sectors. Employing SL-AIV, each sector undergoes a separate analysis, allowing for a potentially more accurate estimation. This sector-based processing makes HO-SIRR more robust in challenging scenarios, such as when multiple reflections arrive simultaneously from different directions, where the single-source assumption of SIRR may fail. By assigning simultaneous reflections to different sectors, HO-SIRR maintains sparsity and avoids overly diffuse rendering.

During the synthesis stage, the directional components from all sectors are panned via VBAP to their respective DOAs and summed. Meanwhile, the diffuse components are first scaled by their sector diffuseness values, re-encoded into the SH domain, decoded to loudspeaker signals using Ambisonics decoding, and finally decorrelated (AmbiDec). This sector-based diffuse rendering allows HO-SIRR to reproduce anisotropic late reverberation, addressing a limitation of the isotropic diffuse model in first-order SIRR.

The binaural variant of HO-SIRR, as detailed by Hold et al. [21], addresses coloration and HRTF resolution in virtual speaker binauralization. This approach integrates HRTFs directly into the synthesis phase of HO-SIRR, rendering directional components according to arrival directions and diffuse components by their sector steering directions. Objective analyses have demonstrated a reduction in coloration compared to the traditional HO-SIRR method.

IV. Experimental Design

A. Evaluation Method

The main goal of this study is to determine which synthetic binaural room impulse response (BRIR) yields an auralization most perceptually similar to that produced by a BRIR recorded with the KU100 dummy head. To evaluate this similarity, we adopted the fidelity attribute, as defined by Zieliński et al. [48]. They described it as the "trueness of reproduction quality to that of the original". The experiment had two dependent variables: (i) spatial fidelity, and (ii) timbral fidelity.

The experiment used the MUSHRA methodology [49], in which participants rated the similarity of test sounds to a reference on a continuous scale from 1.0 ("Extremely different") to 5.0 ("Same"), with intermediate values of 2.0 ("Very different"), 3.0 ("Different"), and 4.0 ("Slightly different"). This scale has been used in analogous subjective studies employing the MUSHRA methodology [9], [27], [50], [51].

The listening tests were conducted in the Applied Psychoacoustics Laboratory (APL) at the University of Huddersfield. Participants listened to the audio stimuli through headphones while seated in the APL's critical listening room. The experiment used fixed head orientations to isolate variables related to the parametric sound field rendering methods, enabling a focused evaluation of their performance in reproducing the characteristics of the critical listening room at specific source positions. The HULTI-GEN Version 2 software provided the test interface [52]. The study was structured around a multifactor design, focusing on two attributes (ATTR): spatial fidelity and timbral fidelity. The evaluation for each attribute was divided into six sessions, each dedicated to a different source position (POSITION). The average session duration was approximately 15 minutes. Furthermore, within each session, there were three specific trials (ITEM), each assessing a distinct type of program material. During each trial, participants rated 10 test conditions (SYSTEM), presented in Table II. Before the experiment, subjects were provided with a detailed instruction sheet. The purpose of the document was to familiarize subjects with the procedure and introduce them to the definitions of spatial and timbral fidelity and methodology.

B. Measurement of Room Impulse Responses

Room impulse response measurements were conducted in the critical listening room of the APL at the University of Huddersfield, which is an ITU-R BS.1116-compliant listening room (6.2m $\times$ 5.6m $\times$ 3.4m; RT 0.25s; NR 12). The loudspeakers used for the measurements were Genelec 8040A, offering a free field frequency response within $\pm$ 2.0dB across a range from 48Hz to 20kHz. The microphone systems used for the measurements were as follows:

TABLE I
TESTED LOUDSPEAKER POSITIONS IN THE ITU-R 4+7+0 LAYOUT

Position	Azimuth	Elevation
Front center	0°	0°
Front left	30°	0°
Side left	90°	0°
Back left	135°	0°
Upper front left	45°	45°
Upper back left	135°	45°

- Neumann KU100 dummy head microphone
- Line Audio OM1 microphone (20Hz to 20kHz, ± 1dB)
- mhAcoustics Eigenmike em32 (referred to as em32 hereon)
- A 6OM1 open microphone array, comprising six omnidirectional Line Audio OM1 microphones. These were arranged in a three-dimensional grid with each microphone pair spaced 100mm apart along the X, Y, and Z axes, closely mimicking the GRAS 50VI intensity probe array as in [9]

BRIRs acquired using the KU100 were used to create reference stimuli for the listening tests. Room impulse responses (RIRs) measured with the em32 were used for rendering stimuli for the SDM and HO-SIRR methods. The particular microphone system was chosen for its 32 capsules, potentially benefiting the SDM utilizing the TDOA-based DOA estimation [9] and enabling 4th order SH encoding, which HO-SIRR leverages for improved rendering via enhanced DOA estimation and anisotropic diffuse modeling [17]. In contrast, an open mic array was used for the SDM conditions, chosen based on studies suggesting its optimal performance in terms of DOA error and perceptual quality [20], [26]. RIRs captured with the dedicated center reference microphone (Line Audio OM1) served as a pressure signal for the SDM conditions.

For the KU100, Line Audio OM1, and 6OM1 array, the Merging Horus audio interface served as the AD/DA converter and microphone preamp. The measurements performed with em32 involved the use of the Eigenmike Microphone Interface Box (EMIB) as the recording device and the Merging Horus as the playback device. To counter potential impulse response distortions due to clock mismatch between devices [53], word clock was used for synchronization, with the Merging Horus device set as master.

The Exponential Sine Sweep (ESS) was used as an excitation signal as described in [53], with the following parameters: a sample rate of 48kHz, a frequency range of 20Hz to 20kHz, a sweep length of 20 seconds, and a fade in/out of 10ms.

The acoustic centers of the microphone systems were positioned at a height of 127.5cm from the floor, aligning with the heights of the acoustic axes of the zero elevation loudspeakers, which form the zero elevation plane of the system. Although our measurements encompassed a complete Dolby 7.1.4 setup, conforming to the 4+7+0 loudspeaker layout as per [4], the study primarily focused on a subset of these configurations, as detailed in Table I. For azimuth angle measurements, defined as the angle relative to the zero azimuth, each loudspeaker at the zero elevation level was placed at a distance of 2.00m (±0.02m) from the center of the microphone array. For the elevation angle measurements, the loudspeakers were placed at a distance of 1.92m (±0.1m) from the center of the microphone array.

C. Test Conditions and Variables

Table II provides a comprehensive list of the systems under test, along with their key components and characteristics, such as microphone arrays, DOA estimators, pressure signals, diffuse renderers, and processing domains.

In this study, SDM conditions were generated using the SDM Toolbox[3], while the PIV-based DOA estimation algorithm was adapted from Zaunschirm et al. [19]. Given BSDM's growing relevance [6], [7], [54], [55], [56], we included two conditions utilizing the available toolbox[4].

The HO-SIRR condition was rendered via its MATLAB implementation[5]. Both the SDM utilizing PIV-based DOA analysis and HO-SIRR conditions employed SHs from the em32 (1st and 4th order, respectively). The SHs were obtained using the EigenUnits plugins (EigenUnit-Encoder) [58].

Settings for SDM, HO-SIRR, and BSDM adhered to the recommended configurations in their respective toolboxes [10], [44], [57], with the exact settings provided in the supplementary material, Section S.III. The exception was the window size, which was standardized at 64 samples across all frameworks, following Amengual Garí et al.'s [20] recommendations . Additionally, post-equalization [12] was disabled in the SDM Toolbox. BSDM and PIV-based methods applied band-limited DOA estimation (200Hz–2400Hz), aligned with the spatial aliasing frequency of the 6OM1 array. BSDM's mixing time was set at 38ms, based on ISM simulations of the auralized room. Conditions with "Omni" appended to their names used a dedicated omnidirectional microphone for spatialization. For conditions without the "Omni" suffix, the spatialized signal was derived from: (i) capsule no. 29 of em32 for SDM-em32; (ii) omnidirectional eigenbeam for SDM-PIV and HO-SIRR.

We ensured that the anchor demonstrated impairments in both spatial and timbral fidelity. Specifically, for azimuthal sources at zero elevation positions, the anchor employed a KU100 BRIR that was rotated by an additional 60 to 90 degrees from the source position being evaluated. For sources at elevated positions, given the limited number of measured positions, a BRIR corresponding to a diametrically opposite position was chosen, 180 degrees from the source position being evaluated. To introduce a timbral fidelity impairment, we applied a 3.5kHz low-pass filter [49].

D. Synthesis of Binaural Room Impulse Responses

To facilitate a subjective experiment, we employed binaural rendering techniques for SDM, BSDM, and HO-SIRR to incorporate the KU100 BRIR as a ground truth reference. Consequently, we employed a dataset of 2702 KU100 head-related impulse responses (HRIRs) sampled on the Lebedev grid [59]. The synthesis process involved rendering virtual loudspeaker

[3]SDM Toolbox (version 1.3001, Updated 22 Apr 2018) [10].
[4]BinauralSDM (version 0.5, commit 965da5c) [44].
[5]HO-SIRR (commit 085f20d) [57].

TABLE II
SYSTEMS UNDER TEST

SYSTEM	Mic Array	DOA Estimator	Pressure signal	Diffuse Renderer	Processing Domain
Anchor	—	—	—	—	—
BSDM-6OM1-Omni	6OM1	TDOA via GCC (Single-source)	Line Audio OM1	RTMod+AP*	Time-domain
BSDM-em32-Omni	em32	TDOA via GCC (Single-source)	Line Audio OM1	RTMod+AP*	Time-domain
SDM-em32	em32	TDOA via GCC (Single-source)	em32 (capsule no. 29)	—	Time-domain
SDM-em32-Omni	em32	TDOA via GCC (Single-source)	Line Audio OM1	—	Time-domain
SDM-6OM1-Omni	6OM1	TDOA via GCC (Single-source)	Line Audio OM1	—	Time-domain
HO-SIRR	em32 (SH)	SL-AIV (Multi-source)	Y_0^0 (B-Format W)	AmbiDec	Time-frequency domain
SDM-PIV	em32 (SH)	PIV (Single-source)	Y_0^0 (B-Format W)	—	Time-domain
SDM-PIV-Omni	em32 (SH)	PIV (Single-source)	Line Audio OM1	—	Time-domain
KU100 (Reference)	—	—	—	—	—

* Please note that RTMod+AP is more of a compensation method rather than an explicit diffuse renderer.

signals at points on the Lebedev grid using the evaluated systems. For SDM, this was achieved using K-nearest neighbor allocation, while HO-SIRR used VBAP. The final stage involved convolving these virtual loudspeaker signals with corresponding HRTFs from the SOFA file [60] and summing them to produce the final BRIRs.

E. Programme Material

Accurate binaural sound field reproduction requires anechoic audio, free from original recording environment effects. This is achieved by convolving the anechoic material with binaural room impulse responses (BRIR), simulating room acoustics. The anechoic samples employed in our study are as follows: "Bongo" (Track 26, 12.78s–20.33s) and "Danish Speech" (Track 9, 0.59s–9.24s) from Bang & Olufsen's "Music for Archimedes" [61], [62], and Handel/Harty's "No.6 Water Music Suite" (Track 9, 0.48s–8.38s) from Denon's anechoic orchestral collection [63].

F. Reproduction Configuration and Calibration

The study employed a reproduction system that included AKG K702 headphones and a Merging Technologies Horus audio interface. These headphones were fitted with an inverse filter originally developed for Lee et al.'s research [64], designed to correct frequency response deviations and replicate the response in a KU100 dummy head's ear. The filter's creation involved placing AKG K702 headphones on a KU100 dummy head and conducting five measurements for each ear in both left/right and right/left positions, then repeating the process with another pair of headphones, resulting in 40 impulse responses in total. The filter design followed the high-pass-regularized least-mean-square (LMS) inversion approach [65], identified as a perceptually sound inversion algorithm [66].

Test stimuli were uniformly normalized to -26 LUFS, and the headphone amp gain was set uniformly throughout the experiment. LAeq loudness for KU100 BRIR at +30° azimuth, measured via miniDSP EARS, averaged 70.7dB (Bongo), 77.5dB (Speech), and 73.5dB (Orchestra). The stimuli employed in the experiment are provided on a public repository.[6]

[6][Online]. Available: https://doi.org/10.5281/zenodo.11165056

TABLE III
RESULTS OF THE AGGREGATED FRIEDMAN TEST FOR DIFFERENT FACTORS AND ATTRIBUTES

ATTR	Factor	Test Statistic (χ^2)	DF	p-value	sig.
Spatial	SYSTEM	101	9	$< .001$	****
Spatial	ITEM	2.43	2	.297	ns
Spatial	POSITION	21.4	5	.007	***
Timbral	SYSTEM	106	9	$< .001$	****
Timbral	ITEM	0.531	2	.767	ns
Timbral	POSITION	11.5	5	.042	*

G. Test Subjects

In total, 14 participants took part in the study. These participants ranged in age from 20 to 47 years. They were a mixture of staff members, postgraduate students, and undergraduate researchers affiliated with the Applied Psychoacoustics Laboratory (APL) at the University of Huddersfield. All participants claimed to have normal hearing abilities and had previously taken part in listening experiments. Prior to the formal listening test, the participants were asked to complete a short 10-minute familiarization phase for each attribute (ATTR).

V. SUBJECTIVE EVALUATION RESULTS

Following ITU guidelines [49], a post-screening process was implemented to identify unreliable assessors. The criterion was rating the hidden reference below 4.5 (90% of the 5.0 maximum score) for over 15% of test items. Ultimately, none of the assessors met these exclusion criteria.

To verify rating normality, data were grouped by SYSTEM, ATTR, ITEM, and POSITION. Ratings for KU100 (Reference) and Anchor consistently showed non-normal distributions. Excluding these, non-normal distributions were found in 6.25% of cases for spatial fidelity and 6.94% for timbral fidelity. Given these results, a non-parametric statistical approach was adopted for subsequent analyses.

Friedman's tests were conducted to examine the main effects of SYSTEM, ITEM, and POSITION on spatial and timbral fidelity attributes. Table III summarizes the detailed statistical outcomes. Results indicate a significant influence of SYSTEM on both attributes, indicating variations in the capability of the rendering methods to replicate spatial and timbral fidelity. Interestingly, the specific program material (ITEM) did not have

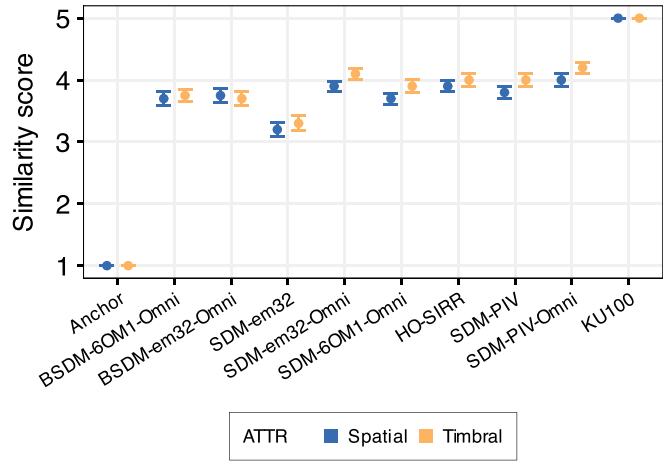


Fig. 1. Subjective evaluation results showing median spatial and timbral fidelity scores on similarity scale (1-5), for various systems (SYSTEM). Scores are compiled from multiple source positions (POSITION) and program materials (ITEM), with 95% non-parametric confidence intervals.

a significant impact on participant ratings, while the source position (POSITION) demonstrated a noticeable effect, especially in spatial fidelity.

The aggregated results across ITEMs and POSITIONs provide an overview of system performance, as shown in Fig. 1. Based on the Friedman test findings, we conducted further analyses on position-based performance (Fig. 2) and potential subtle influences of program material (Fig. 3). Differences between rendering systems were analyzed using the Wilcoxon signed-rank test with Holm-Bonferroni correction [67]. Effect sizes (r) were calculated as described in [68], converting p-values to z-scores and then to r, based on total observations (N).

Fig. 1 illustrates spatial and timbral fidelity ratings across systems. Anchor consistently scored at the lower end of the scale with a median of 1.0, indicating "Extremely Different". In contrast, KU100 achieved a median score of 5.0 ("Same"), implying that subjects had no problem identifying the hidden reference. The performance of the evaluated systems, from BSDM-6OM1-Omni to SDM-PIV-Omni, varied, with spatial fidelity ratings mostly between 3 and 4 and timbral fidelity ratings between 3 and 4.2 on the similarity scale.

Looking more closely at spatial fidelity, SDM-em32-Omni, HO-SIRR, and SDM-PIV-Omni displayed comparable performance (medians between 3.9 and 4). However, SDM-PIV, with a median of 3.8, was slightly inferior to both SDM-em32-Omni and HO-SIRR. SDM-6OM1-Omni trailed behind these three systems. Additionally, BSDM-6OM1-Omni, BSDM-em32-Omni, and SDM-6OM1-Omni were grouped closely together (medians between 3.7 and 3.8), though BSDM-em32-Omni was outperformed by both HO-SIRR and SDM-PIV. Notably, SDM-em32 performance was significantly lower than the rest (median 3.2).

In terms of timbral fidelity, SDM-em32-Omni and SDM-PIV-Omni were almost indistinguishable (medians between 4.1 and 4.2), with SDM-em32-Omni outperforming HO-SIRR and SDM-PIV. SDM-6OM1-Omni, HO-SIRR, and SDM-PIV were on par with each other (medians between 3.9 and 4), whereas BSDM-6OM1-Omni and BSDM-em32-Omni, while similar to each other (medians 3.7), lagged behind the preceding group, and even more so when compared to SDM-em32-Omni and SDM-PIV-Omni. Again, SDM-em32 scored substantially lower than BSDM-6OM1-Omni and BSDM-em32-Omni.

A. *Dedicated Center Omnidirectional Microphone*

The impact of using a dedicated center omnidirectional microphone can be analyzed by comparing SDM-em32-Omni and SDM-em32—employing TDOA-based DOA—as well as SDM-PIV-Omni and SDM-PIV—utilizing PIV-based DOA. These conditions used SRIR from em32 (and obtained SHs) with and without a center omnidirectional microphone.

The dedicated omnidirectional microphone (SDM-em32-Omni) notably enhanced the em32 array's performance in the context of SDM's TDOA-based DOA variant (SDM-em32), with significant improvements in aggregated results for both spatial and timbral fidelity, exhibiting a large effect size ($p < .001$, $r > .70$). The same trend was also observed for individual positions and stimuli, however, for certain positions with medium effect size ($r = .40$), for instance, +45°, +45° (spatial fidelity, $p = .005$) and +135°, +45° (timbral fidelity, $p = .004$).

Improvements were also noticeable with a dedicated center microphone in SDM's PIV-based DOA estimation variant, but they were less pronounced. Compared to SDM-PIV's zeroth-order eigenbeam, SDM-PIV-Omni's dedicated omnidirectional microphone showed significant enhancement in spatial ($p = .009$, $r = .16$) and timbral fidelity ($p < .001$, $r = .29$), both with small effect sizes. Notably, for the +135° source position, the improvement in timbral fidelity was significant with a large effect size ($p < .001$, $r = .61$). Across most source types, the improvement for both attributes was observed with a small effect size ($r < .26$), except for Bongo and Speech in spatial fidelity, where the improvement was not significant.

B. *Microphone Array and DOA Estimation Method*

This research evaluated TDOA and PIV-based DOA estimation methods in SDM for spatial data capture and auralization. It focused on TDOA-based DOA using em32 and 6OM1 arrays, represented by SDM-em32-Omni and SDM-6OM1-Omni, and PIV-based DOA with first-order SHs, denoted as SDM-PIV-Omni, all utilizing a central pressure signal from the array.

As depicted in Fig. 1, our results indicated that SDM-em32-Omni, which relies on TDOA-based DOA estimation, exhibited performance on par with SDM-PIV-Omni—using a PIV-based DOA estimation—when identical omnidirectional pressure signals were used. The comparison revealed no substantial differences in either spatial or timbral fidelity. Conversely, SDM-6OM1-Omni—also based on TDOA but incorporating an array of only six microphones—was found to perform significantly worse than SDM-em32-Omni with a small effect size ($p < .001$, $r < .26$) and SDM-PIV-Omni with a medium effect size ($p < .001$, $.30 < r < .36$) for both spatial and timbral fidelity.

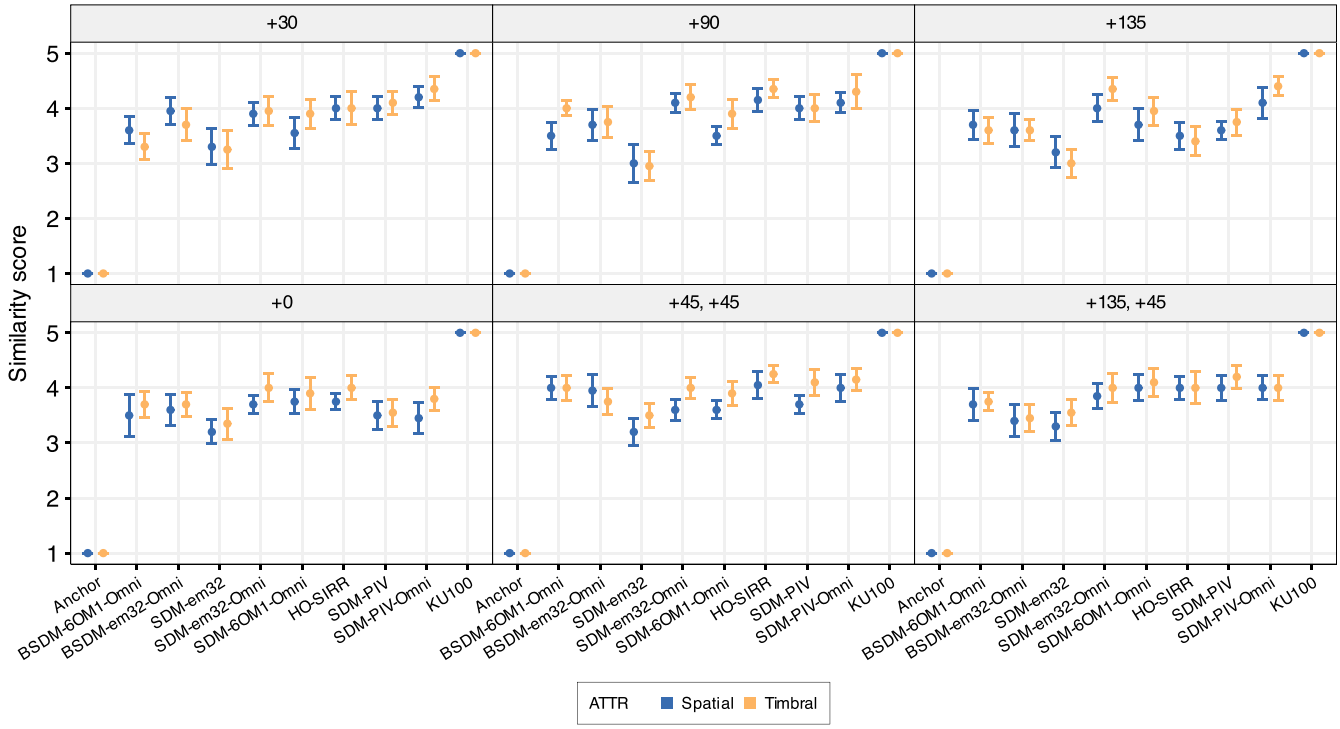


Fig. 2. Subjective evaluation results showing median for spatial and timbral fidelity scores on similarity scale (1-5), across different systems (SYSTEM) and source positions (POSITION), aggregated over all program materials (ITEM). The graph includes 95% non-parametric confidence intervals.

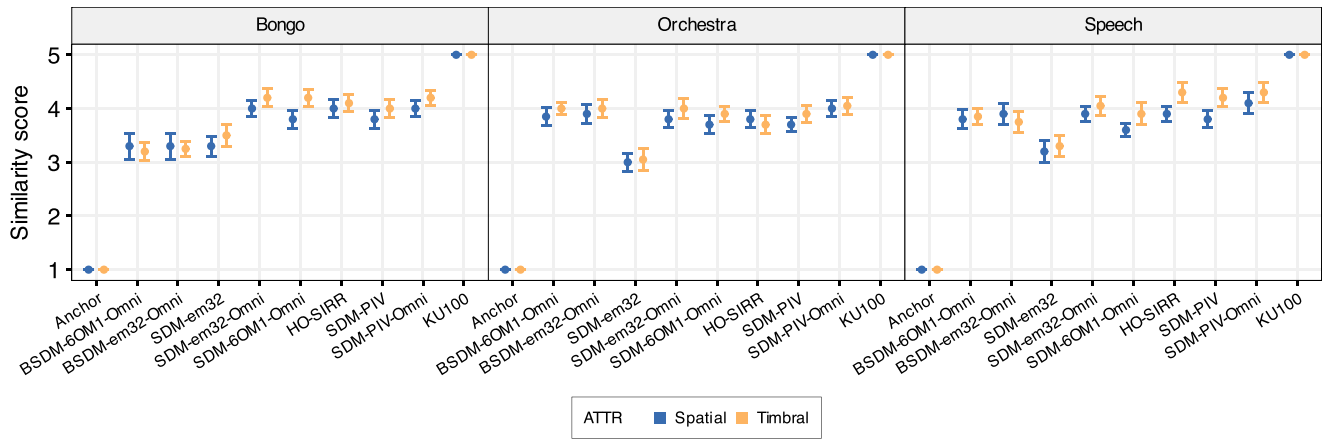


Fig. 3. Subjective evaluation results showing median for spatial and timbral fidelity scores on similarity scale (1-5), across different systems (SYSTEM) and program materials (ITEM), aggregated over all source positions (POSITION). The graph includes 95% non-parametric confidence intervals.

Upon examining the specific source positions (Fig. 2), we observed a similar pattern, particularly pronounced for source positions at +30° and +90°. For spatial fidelity, SDM-6OM1-Omni demonstrated significant deviation from SDM-em32-Omni ($p = .007, r = .41$ at +30°; $p < .001, r = .61$ at +90°) and SDM-PIV-Omni ($p < .001$, $r = .66$ at +30°; $p < .001$, $r = .69$ at +90°). Effect sizes ranged from medium to large. On the other hand, SDM-PIV-Omni outperformed SDM-em32-Omni in spatial fidelity at +30° ($p = .013$, $r = .38$) and +45°, +45° ($p = .003$, $r = .45$) source positions with medium effect sizes. Regarding timbral fidelity, SDM-em32-Omni outperformed SDM-6OM1-Omni at +90° ($p = .002$, $r = .47$) and +135° ($p < .001$, $r = .50$) with medium to large effect sizes. SDM-PIV-Omni also showed better timbral fidelity than SDM-6OM1-Omni at multiple positions: +30° and +90° ($p = .011$, $r = .39$, medium effect size), +135° ($p < .001$, $r = .54$, large effect size), and +45°, +45° ($p = .006$, $r = .42$, medium effect size).

Stimulus-dependent analysis revealed minimal differences between SDM-em32-Omni and SDM-PIV-Omni, with SDM-PIV-Omni showing slightly better spatial fidelity for Orchestra stimulus, with a small effect size ($p = .045$, $r = .22$). For Speech stimulus, SDM-PIV-Omni also showed greater timbral fidelity than SDM-em32-Omni, again with a small effect size ($p = .028$, $r = .24$). In contrast, SDM-6OM1-Omni

underperformed compared to SDM-em32-Omni (spatial fidelity: $p = .005$, $r = .30$; timbral fidelity: $p = .001$, $r = .35$) and SDM-PIV-Omni (spatial fidelity: $p = .003, r = .32$; timbral fidelity: $p < .001, r = .57$) for Speech stimulus with medium to large effect sizes. For Orchestra stimulus, SDM-PIV-Omni outperformed SDM-6OM1-Omni in both spatial fidelity ($p < .001$, $r = .36$, medium effect size) and timbral fidelity ($p = .022$, $r = .25$, small effect size).

C. Spatial Reproduction Systems

This section examines the impact of different spatial reproduction frameworks on spatial and timbral fidelity relative to the reference. We compare several SDM variants (SDM-em32, SDM-em32-Omni, SDM-6OM1-Omni, SDM-PIV, SDM-PIV-Omni), HO-SIRR, and BSDM optimizations (BSDM-6OM1-Omni, BSDM-em32-Omni).

The results showed no significant difference in spatial fidelity between SDM-6OM1-Omni and BSDM-6OM1-Omni. However, SDM-6OM1-Omni slightly outperformed BSDM-6OM1-Omni in timbral fidelity ($p = .001$, $r = .20$). Notably, BSDM optimizations often led to degradation ($p < .001$) in both spatial ($r = .25$) and timbral fidelity ($r = .49$) compared to their SDM counterparts, as evidenced by the comparison between BSDM-em32-Omni and SDM-em32-Omni.

Analysis of performance across different source positions revealed that BSDM-6OM1-Omni generally matched SDM-6OM1-Omni in fidelity, with exceptions at specific angles. At +30°, SDM-6OM1-Omni significantly outperformed BSDM-6OM1-Omni in timbral fidelity ($p < .001, r = .54$), and BSDM-6OM1-Omni underperformed compared to systems SDM-em32-Omni to SDM-PIV-Omni with a large effect size ($p < .001$, $r > .52$) in the same aspect. However, no significant difference was observed in spatial fidelity at this angle. At +90°, HO-SIRR, SDM-PIV, and SDM-PIV-Omni outperformed BSDM-6OM1-Omni in spatial fidelity ($p < .001, r > .50$). Additionally, at the +135°, +45° position, BSDM-6OM1-Omni underperformed in terms of both spatial ($p = .008, r = .40$) and timbral ($p = .016$, $r = .37$) fidelity compared to SDM-6OM1-Omni.

The comparison between BSDM-em32-Omni and SDM-em32-Omni showed that both systems performed similarly across most positions, except at +90° and +135°, +45°. At these angles, SDM-em32-Omni consistently outperformed BSDM-em32-Omni. For spatial fidelity, SDM-em32-Omni showed superior performance at +90° ($p < .001$, $r = .51$, large effect size) and at +135°, +45° ($p = .004$, $r = .44$, medium effect size). Regarding timbral fidelity, SDM-em32-Omni outperformed BSDM-em32-Omni at +90° ($p < .001$, $r = .64$, large effect size), +135° ($p = .001$, $r = .53$, large effect size), and +135°, +45° ($p = .002$, $r = .47$, medium effect size).

Analysis of individual stimuli showed BSDM-6OM1-Omni and BSDM-em32-Omni matching SDM and HO-SIRR systems in spatial fidelity for Orchestra and Speech stimuli, but significantly underperforming for Bongo stimulus ($p < .001$, $r > .5$, large effect size). Timbral fidelity ratings showed a similar trend, with BSDM variants occasionally underperforming SDM-em32-Omni, SDM-6OM1-Omni, SDM-PIV, and SDM-PIV-Omni for Orchestra and Speech stimuli (effect sizes ranging from small to large).

The superior performance of SDM-em32-Omni and SDM-PIV-Omni over SDM-6OM1-Omni suggests the impact of the microphone array used. Comparing SDM variants with HO-SIRR revealed no significant difference in spatial fidelity among HO-SIRR, SDM-em32-Omni, SDM-PIV, and SDM-PIV-Omni. However, HO-SIRR outperformed SDM-6OM1-Omni in spatial fidelity ($p = .009$, $r = .16$), while underperforming compared to SDM-em32-Omni ($p = .018$, $r = .15$) and SDM-PIV-Omni ($p < .001, r = .27$) in timbral fidelity, all with small effect sizes. Notably, HO-SIRR and SDM-PIV showed similar performance in both aspects, potentially due to their shared use of the zeroth-order eigenbeam as the pressure signal.

The results for individual source positions implied that HO-SIRR performed similarly to SDM-em32-Omni and SDM-PIV-Omni, with some exceptions. HO-SIRR underperformed at +135° for timbral fidelity ($p < .001$, $r > .66$, large effect size), but outperformed SDM-em32-Omni at +45°, +45° for spatial fidelity ($p = .024$, $r = .34$, medium effect size). Across different program materials, HO-SIRR and SDM-PIV showed no significant differences. However, for the Orchestra stimulus, HO-SIRR was outperformed in timbral fidelity by SDM-em32-Omni ($p = .004$, $r = .31$) and SDM-PIV-Omni ($p < .001$, $r > .37$), both with medium effect sizes.

D. Principal Component Analysis

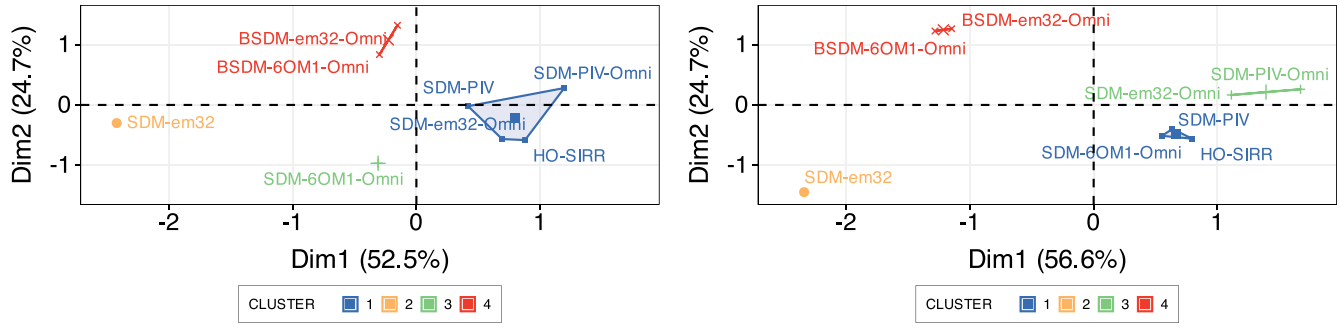
In the spatial analysis and auralization of a critical listening room, SDM and HO-SIRR systems exhibited similar performance, particularly in terms of spatial fidelity. Systems such as SDM-em32-Omni, HO-SIRR, SDM-PIV-Omni, and SDM-PIV demonstrated comparable results, though with some variations. With respect to timbral fidelity, SDM-em32-Omni and SDM-PIV-Omni, which utilize TDOA-based analysis with em32 and PIV-based analysis with a dedicated pressure signal, closely aligned with the reference.

This alignment is supported by Principal Component Analysis (PCA), as shown in Fig. 4. The PCA clusters these systems together based on their median spatial and timbral fidelity scores across all positions and stimuli. The analysis revealed that the first two principal components accounted for about 77% of the variance in spatial fidelity and 80% in timbral fidelity.

E. Correlation Between Spatial and Timbral Fidelity

To explore the link between spatial and timbral fidelity, we calculated Spearman's correlation based on median scores across all positions and stimuli, as depicted in Fig. 5.

The weakest correlations between spatial and timbral fidelity were observed in SDM-em32-Omni ($\rho = .41$), SDM-em32, and SDM-6OM1-Omni ($\rho = .45$), followed by SDM-PIV-Omni ($\rho = .52$). A significant increase in correlation was noted for HO-SIRR ($\rho = .59$) and SDM-PIV ($\rho = .60$), with BSDM-em32-Omni ($\rho = .72$) and BSDM-6OM1-Omni ($\rho = .77$) exhibiting the strongest positive correlations.



(a) Spatial Fidelity (b) Timbral Fidelity

Fig. 4. Principal Component Analysis (PCA) of spatial and timbral fidelity scores for evaluated systems (SYSTEM), considering the median spatial and timbral fidelity scores across different program materials (ITEM) and source positions (SYSTEM). Systems are clustered in a two-dimensional space by the first two principal components, highlighting the similarities in their fidelity scores.

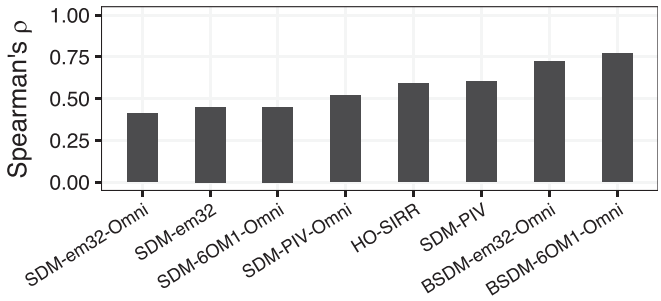


Fig. 5. Correlation between median scores of spatial and timbral fidelity across all source positions and stimuli.

The strong correlation in BSDM-rendered conditions indicates a uniform impact of artifacts on both spatial and timbral fidelity, possibly due to reverb equalization effects. Conversely, the pronounced correlation between spatial and timbral fidelity in HO-SIRR and SDM-PIV potentially stems from using a zeroth-order SH as the pressure signal. As demonstrated in the supplementary material (Sections S.I and S.II), the zeroth-order SH from em32 suffers from spatial aliasing beyond 8.5kHz, deviating from the ideal directional pattern and frequency response. This directly impacts timbre but also could affect spectral cues in the HRTFs during BRIR synthesis. If a sound source lacks spectral details, the direction-dependent filtering of the external ear cannot impose these cues, consequently influencing spatial perception [69], [70], [71].

VI. Objective Metrics

The present study considered five objective metrics: Interaural Level Difference (ILD), Interaural Time Difference (ITD), reverberation time (RT_{60}), Interaural Cross-Correlation Coefficient (early, $IACC_{E3}$; late, $IACC_{L3}$), and Energy Difference (ED). To evaluate error relative to the reference, the mean absolute error (MAE) measures error magnitude, while the mean signed difference (MSD) identifies systematic biases, indicating the extent and direction of performance deviations.

Each BRIR was energy-normalized to allow a fair comparison. This involved identifying the onset of direct sound, indicated by the earliest sound arrival in either channel, and calculating the Root Mean Square (RMS) value over a 2.5ms segment starting from this point [72]. This RMS value, representing the direct sound's energy, was then used to normalize the entire BRIR, thus minimizing gain differences and establishing a consistent baseline for objective analysis.

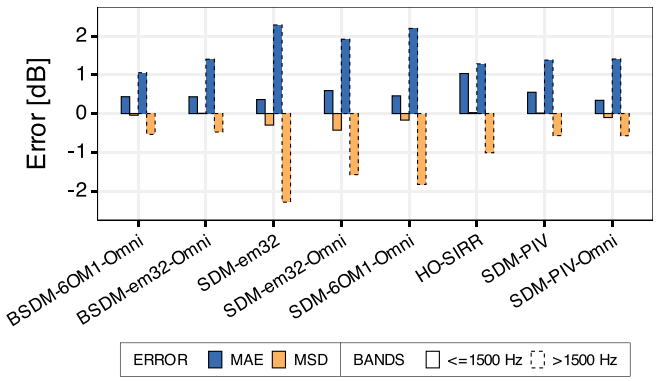
A. Interaural Level Difference (ILD)

The ILD is a major cue for horizontal sound localization, accentuated by head shadowing when sources are off-center, with the Just Noticeable Difference (JND) ranging from 1 to 2 dB [75]. Accounting for the precedence effect, ILD was derived from the first 2.5ms of BRIRs over 39 equivalent rectangular bandwidth (ERB) bands and averaged

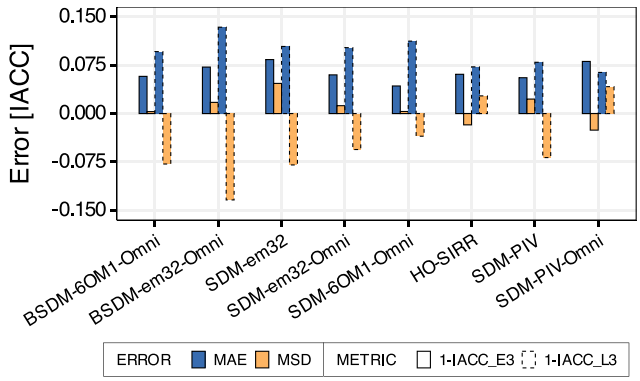
$$\mathrm{ILD}_{\mathrm{avg}} = \frac{1}{N}\sum_{f=1}^{N} 20\log_{10}\left(\frac{\hat{y}_L(f)}{\hat{y}_R(f)}\right). \tag{1}$$

Here, $\mathrm{ILD}_{\mathrm{avg}}$ is the average ILD over N ERB bands, with $\hat{y}_L(f)$ and $\hat{y}_R(f)$ as the RMS values for left and right channels at each frequency band f.

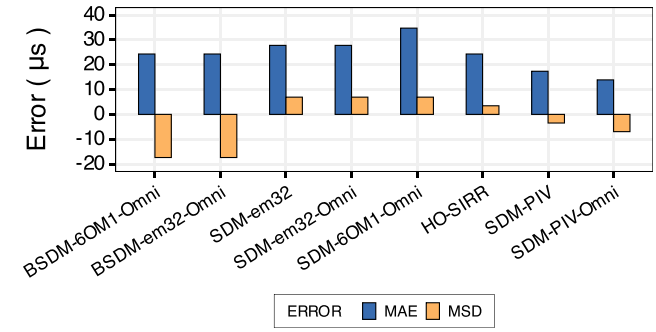
Fig. 6(a) illustrates the MAE and MSD for ILD across evaluated systems, categorized by the ERB bands above and below 1500Hz used for averaging. At higher frequencies, SDM-em32, SDM-em32-Omni, and SDM-6OM1-Omni showed notable deviation with high MAEs and negative MSDs, underscoring a tendency to underestimate ILDs beyond the established JND [75]. HO-SIRR also tended to underestimate ILD, but to a lesser extent. In contrast, lower frequencies exhibited reduced MAEs for most systems, except HO-SIRR, and MSDs near zero, indicating minimal bias except for the consistent underestimation by SDM-em32, SDM-em32-Omni, and SDM-6OM1-Omni. Overall, all systems faced more difficulty with higher-frequency ILDs, implying challenges in capturing spatial cues accurately. However, performance improved at frequencies below 1500Hz, with SDM-em32 being the exception.



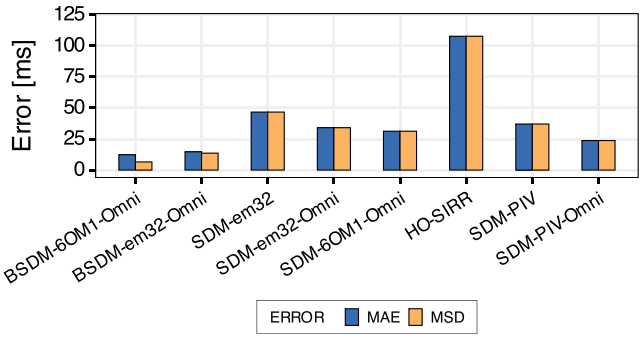
(a) ILD computed from BRIRs across 39 ERB bands, averaged above and below 1.5kHz as detailed in [73].



(b) Interaural cross-correlation coefficients: Early (1-IACC$_{E3}$) and Late (1-IACC$_{L3}$).



(c) ITD calculated using maximum interaural cross-correlation (MaxIACCr) method.



(d) Mid-frequency reverberation time T30$_{mid}$ as defined in [74].

Fig. 6. Objective metrics presented using mean absolute error (MAE) and mean signed difference (MSD) using the reference KU100 as the ground truth.

B. Interaural Time Difference (ITD)

The ITD is the second of two auditory cues critical for lateral sound localization, with JND of about 40μs for frontal sources and approximately 100μs for lateral sources [76]. The ITD can extend up to 700μs for azimuth angles up to 90 degrees. ITD estimation is based on the peak time lag of the interaural cross-correlation function

$$\text{ITD} = \arg \max_{-1\ \text{ms}<\tau<1\ \text{ms}} (|\text{IACF}(\tau)|). \quad (2)$$

Fig. 6(c) demonstrates ITD accuracy with MAE and MSD values, indicating that BSDM-6OM1-Omni and BSDM-em32-Omni consistently underestimate ITD, whereas SDM-em32, SDM-em32-Omni, and SDM-6OM1-Omni overestimate, particularly SDM-6OM1-Omni with the highest MAE of 34.7μs. HO-SIRR has minor overestimations, while SDM-PIV and SDM-PIV-Omni maintain the highest accuracy with slight underestimation tendencies. Given that the JND for ITD ranges from 40μs frontally to 100μs laterally [76], most systems' deviations fall within perceptually insignificant limits. Notably, there is an inconsistency in MSD and MAE magnitudes and a suggested ILD-ITD trade-off in systems SDM-em32 through HO-SIRR.

C. Reverberberation Time (RT_{60})

The RT$_{60}$ was calculated in octave bands in accordance with the standards outlined in [74]. The value presented was derived based on a 30dB evaluation range and subsequently averaged across the 500Hz and 1kHz octave bands, resulting in T30$_{\text{mid}}$. The JND is established at 5% of the RT$_{60}$ [74]. With a 0.25-second RT$_{60}$ in the auralized room, the JND in the present scenario is roughly 12.5ms.

Fig. 6(d) shows that all systems tend to overestimate RT$_{60}$, as indicated by positive MSD values. Most systems' errors exceed the JND of 12.5ms, which may affect perception, except for BSDM-6OM1-Omni and BSDM-em32-Omni, which maintained minimal perceptible errors with MAEs of 12.4ms and 14.7ms. These results align with Amengual Garí et al.'s findings, confirming that RTMod maintains synthesized BRIR's RT$_{60}$ within the JND [20].

On the other hand, HO-SIRR, with an MAE of 108ms, significantly overshoots the JND, hinting at a noticeable impact on spatial fidelity. The observed deviations may be due to the default decorrelation filter lengths in the HO-SIRR toolbox [57], which range from 200ms at low frequencies to 40ms at high frequencies. HO-SIRR was proposed and evaluated for larger, more reverberant spaces [17]. These filter lengths may therefore be too long for the smaller room in this study.

D. Interaural Cross Correlation Coefficient (IACC)

The IACC is a commonly used metric for assessing spatial impression (SI) in concert halls, with a JND of 0.075 [74]. SI can be further divided into Apparent Source Width (ASW) and

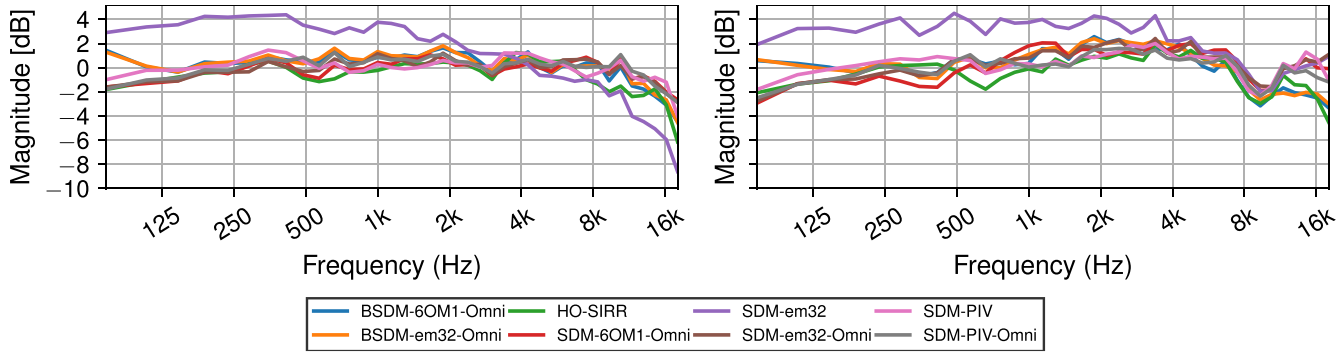


Fig. 7. The energy difference (ED) in decibels for each system under test (SYSTEM). Calculated for 39 ERB bands and averaged across all source positions (POSITION). The left plot displays the ED for the ipsilateral ear, while the right plot shows the ED for the contralateral ear. In both plots, the system's output was compared to the reference condition. Positive ED values indicate the test system has more energy than the reference condition, while negative values indicate less energy. Note that RMS has been normalized before comparison, as explained in Section VI.

Listener Envelopment (LEV). ASW and LEV are best estimated by calculating the IACC in octave bands over early (0–80ms) and late (80–750ms) periods, respectively, and then averaging the IACC values across the 500Hz, 1kHz, and 2kHz octave bands, resulting in the IACC_{E3} and IACC_{L3} metrics [77]. In the present study, we calculated 1-IACC_{E3} and 1-IACC_{L3}, which are positively correlated with ASW and LEV. The IACC is defined as

$$\text{IACC} = \max_{-1\text{ ms}<\tau<1\text{ ms}} |\text{IACF}(\tau)|. \tag{3}$$

Fig. 6(b) presents MAE and MSD for systems using 1-IACC_{E3} and 1-IACC_{L3}. Evaluated systems vary in accuracy, with 1-IACC_{E3} MAE ranging from 0.0427 to 0.0836 and 1-IACC_{L3} MAE from 0.0637 to 0.134. Errors for 1-IACC_{E3} mostly fall within the JND of 0.075, implying that the errors are unlikely to be perceptible. Low absolute MSDs suggest non-systematic errors, except for SDM-em32 and SDM-PIV-Omni, which show a bias in 1-IACC_{E3} estimates, potentially affecting ASW.

In contrast, MAEs for 1-IACC_{L3} surpass the JND of 0.075 for all but HO-SIRR and SDM-PIV-Omni, with several systems showing negative MSDs, hinting at consistent underestimation and possible LEV impact. HO-SIRR and SDM-PIV-Omni, however, remain within JND bounds for both MAE and MSD, suggesting minimal LEV alteration.

E. Energy Difference (ED)

The ED metric compares the energy levels between test and reference conditions across 39 ERB bands from 50Hz to 20 kHz [78]. The ED is calculated as the logarithmic ratio of the energy for a test condition relative to the reference

$$ED^{(l,r)}(f_c) = 10\log_{10}\left(\frac{\int |\mathcal{G}_{f_c}\{p^{(l,r)}(t)\}|^2 dt}{\int |\mathcal{G}_{f_c}\{p^{(l,r)}_{ref}(t)\}|^2 dt}\right), \tag{4}$$

where $ED^{(l,r)}(f_c)$ is the ED for the left (l) and right (r) ears at the center frequency f_c, $p^{(l,r)}(t)$ and $p^{(l,r)}_{ref}(t)$ are the pressure signals for the test and reference conditions, respectively, and $\mathcal{G}_{f_c}\{\cdot\}$ denotes the Gammatone filter with center frequency f_c. Fig. 7 presents the ED across ERBs for each system, split into ipsilateral and contralateral ears. Table IV lists the mean absolute ED (MAED) across ERBs per system.

TABLE IV
MEAN ABSOLUTE ED (MAED) ACROSS 39 ERB BANDS

SYSTEM	Ipsilateral ear	Contralateral ear
BSDM-6OM1-Omni	0.93 dB	1.28 dB
BSDM-em32-Omni	0.91 dB	1.30 dB
HO-SIRR	0.99 dB	1.11 dB
SDM-6OM1-Omni	0.64 dB	1.20 dB
SDM-em32	2.85 dB	2.63 dB
SDM-em32-Omni	0.66 dB	1.08 dB
SDM-PIV	0.63 dB	0.86 dB
SDM-PIV-Omni	0.67 dB	0.83 dB

The most notable observation was the high ED of SDM-em32 up to 2kHz, likely due to the spherical baffle's impact on the frequency-dependent directivity of the pressure signal from capsule no. 29. This directional bias resulted in strong high-frequency attenuation and spectral imbalance, as demonstrated in the octave band analysis of pressure signals in the supplementary material (Section S.II). For the ipsilateral ear, significant deviations and high-frequency loss above 8kHz are also present. Subjective results aligned with ED, indicating strong spatial and timbral fidelity degradation for SDM-em32, consistent with studies on the spectrum's impact on spaciousness and spatial fidelity [48], [79]. For the contralateral ear, a pronounced dip around 9kHz was observed across all systems, possibly due to spectral notches [69] in the HRTFs employed for synthesis, which were absent in the reference condition. This may have originated from KNN mapping or VBAP panning when the exact HRTF was unavailable, or errors in DOA estimation leading to the incorrect HRTF being chosen.

Most systems showed ED within ± 1dB from 125Hz to 10kHz for the ipsilateral ear, with a significant drop above 10kHz. The contralateral ear followed a similar trend. HO-SIRR and SDM-6OM1 showed dips reaching -2dB at 500Hz–1kHz and 250–500Hz, respectively. Between 1–6kHz, systems generally exhibited 1–2dB ED. Above 10kHz, most maintained 0–1dB ED, while HO-SIRR and BSDM systems showed around -2dB ED.

More generalized observations can be made from the MAED values in Table IV. SDM-PIV and SDM-PIV-Omni exhibited the lowest MAED for both ipsilateral and contralateral ears. SDM-em32-Omni and SDM-6OM1-Omni also showed low MAED for the ipsilateral ear but higher values for the contralateral ear. BSDM-6OM1-Omni, BSDM-em32-Omni, and HO-SIRR had higher MAED for both ears, with HO-SIRR's contralateral ear ED being about 0.2dB lower than the BSDM conditions. Given the JNDs of 0.2–0.3dB for 1kHz tones and 0.5dB for broadband noise at 70–80dB SPL [80], the observed ED values are likely perceptible, as observed in the subjective results.

VII. Discussion

McCormack et al. [17] found that post-equalized SDM using PIVs underperformed compared to SIRR in simulated environments with ideal SH input. In contrast, Zaunschirm et al. [18] observed improved SDM performance in real-world settings when PIV-based DOA estimation incorporated band-limitation. Zaunschirm et al. [18] did not specify applying post-equalization for SDM conditions using KNN rendering. The differences between the two studies may be due to the use of post-equalization in KNN-rendered SDM or different experimental designs: real-world measurements [18] versus simulated acoustic environments [17].

Our study found comparable spatial fidelity for HO-SIRR and SDM using TDOA and PIVs for DOA estimation. Although HO-SIRR overestimated RT_{60} by more than 100ms, potentially due to decorrelation filter lengths, subjective evaluation did not reflect this. The effectiveness of SDM, compared to more complex methods such as HO-SIRR, in rendering dry listening rooms suggests a good match between SDM's sound field model and the acoustic characteristics of these spaces, which feature prominent early reflections and shorter, quieter diffuse tails. While certain test systems were rated as "Slightly Different" on the scale in terms of spatial and timbral fidelity, no evaluated system was indistinguishable from the dummy head reference, aligning with the presented objective metrics and previous research [26]. This study is unique in using real-world measurements with a comprehensive set of conditions, offering a realistic context relevant to practical room acoustics applications, contrasting with previous studies' simulated conditions.

The performance of SDM with optimizations for binaural rendering (BSDM) depends on the stimulus type. The Bongo stimulus adversely impacted spatial and timbral fidelity, suggesting that BSDM's equalization quality might be compromised by suboptimal RT_{60} estimation in low-reverb environments [54]. On the other hand, overall results showed that the effect size was small and medium for spatial and timbral fidelity, respectively, indicating that these artifacts could have a more significant impact on timbre, which is also demonstrated by the MAED metric. While objective measurements confirmed RTmod's efficacy, showing the RT_{60}'s MAE within the JND range, the observed poor performance may be linked to artifacts related to RTmod, appearing in transient sounds [20]. BSDM conditions also significantly underestimated 1-$IACC_{L3}$, a key metric for listener envelopment [77]. This underestimation surpassed the established JND [74], suggesting a potential reduction in listener envelopment relative to the reference. These observations, combined with the performance of SDM-PIV in the present and McCormack et al.'s study [17] suggest that post-processing algorithms for SDM should be applied with care, as they may introduce artifacts that could affect spatial and timbral fidelity.

In a listening test comparing binaural renderings to a real loudspeaker's sound, Amengual Garí et al. found that BSDM-rendered auralizations were comparable in plausibility to real loudspeakers [20]. In contrast, our study evaluated BSDM's spatial and timbral fidelity against a ground truth binaural reference, revealing lower fidelity than standard SDM. This finding, along with our pilot study using DOA enforcement for direct sound and band-limited DOA estimation [28], suggests that incorporating core BSDM optimizations into SDM could improve robustness with regard to spatial and timbral fidelity when compared to the ground truth, especially when contrasted with the complete BSDM utilizing RTmod+AP optimization.

Both TDOA and PIVs can be equally effective DOA estimators in SDM when a sufficient number of microphones is utilized for TDOA estimation. In band-limited PIV-based DOA estimation, the condition utilizing an omnidirectional pressure signal demonstrates superior performance compared to the TDOA-based algorithm with a six-microphone array. Previous studies have used different configurations for DOA estimation using the PIV method in SDM. Ahrens [26] conducted a study using PIV without band limitation and found that this SDM variant resulted in larger perceived differences in auralizations compared to other array geometries, using a dummy head reference as the baseline. On the other hand, Zaunschirm et al. [18] applied band limitation to PIV-based DOA estimation, using an effective frequency range from 200Hz to just below the spatial aliasing frequency of the microphone array. Their SDM results closely matched the binaural reference in terms of image width, distance, and diffuseness. Our study aligned with these findings, reinforcing the importance of band limitation in PIV-based DOA estimation for optimal spatial analysis and synthesis when using SHs obtained from real microphone arrays.

We did not expect SDM-em32-Omni to match the performance of SDM-PIV-Omni given the complexity of the TDOA estimation using SMAs and the rigid sphere design of em32 [29]. Although the larger number of microphones in em32 suggests a potential for improved TDOA estimation, objective measurements indicate higher errors in ILD, ITD, and RT_{60}, aligning em32 more with SDM-6OM1 Omni rather than SDM-PIV-Omni. Previous applications of SDM employed small microphone separations with the smallest inter-mic distance of 17.7mm [12]. The 26.5mm inter-mic distance in em32 may be suitable for TDOA estimation, but optimizing the TDOA model for rigid sphere microphones could potentially improve accuracy [81].

SDM-em32-Omni with 32 sensors outperforms the SDM-6OM1-Omni with six sensors. In contrast, Ahrens' study using different array configurations, including a 12-microphone array, found that a six-microphone array, especially with radii of 50mm and 100mm, produced the smallest perceptual differences to the dummy head reference [26]. The superior performance of the

em32 in our study may be attributed to the larger number of microphones in the em32 aiding in localization [9], as well as a higher spatial aliasing frequency, which may positively impact the TDOA estimation [30].

Using a dedicated omnidirectional microphone with the em32 array improves the performance of SDM utilizing TDOA-based DOA estimation, as shown by ILD, IACC, and ED metrics. The em32's top capsule, when used as a pressure signal, has a complex directivity pattern, likely causing ILD errors due to ED inaccuracies. While SDM-PIV-Omni shows improvement over SDM-PIV (using zeroth-order SH), both show minimal MAED from the reference. Subjective evaluation revealed significant improvements in spatial and timbral fidelity, especially for the +135° source position. This may be due to spatial aliasing in the zeroth-order SH beyond 8.5kHz, impacting the directional pattern and frequency response, primarily affecting the +135° position, but not evident in the mean ED across all positions. Octave band energy analysis of pressure signals revealed close matching between dedicated omnidirectional microphone and zeroth-order eigenbeam signals up to 8.5kHz, diverging above, potentially due to spatial aliasing (supplementary material, Section S.II.B). Present findings contrast with Amengual Garí et al. [27], who found no significant difference between dedicated pressure signal and zeroth-order SH in SDM-PIV. Similarities in spatial and timbral fidelity between HO-SIRR and SDM-PIV, along with the improvements observed between SDM-PIV and SDM-PIV-Omni, suggest that adding a dedicated pressure microphone may improve SIRR-based frameworks, as noted in [8].

Pressure signals directly affect timbre and spatial fidelity through cues like ILD. Previous studies demonstrated that spectral imbalance can affect spaciousness, source width, and spatial fidelity [48], [79], [82]. Additionally, significant high-frequency loss may impair localization and front-back discrimination by limiting the extraction of crucial spectral cues [69], [70], [71]. This could explain the strong correlation observed between spatial and timbral fidelity scores in certain conditions, where both attributes were rated similarly. Previous subjective studies have also found close relations in spatial and timbral fidelity patterns [64], [83]. This, combined with observations by Zieliński et al. [48] and the present findings, suggests an interaction between spatial and timbral fidelity. Impairments in one attribute can affect the other, presenting an important consideration for experimental design and an avenue for further research.

The present study was limited to fixed head orientations. While this approach allowed for controlled evaluation at specific static source positions, it diverged from natural listening scenarios. Blauert [75] emphasizes that listeners use variations in binaural signals during head movement as localization cues. The absence of these dynamic cues can contribute to localization errors, front-back confusion, and poor externalization [75]. Similar observations were made by Begault et al. [84], although the authors implied that early reflections could be a dominant factor influencing externalization. The static head orientation in our study did not allow for potential differences between test conditions and the reference that may be perceived with head movement. Future studies that evaluate the methods in the context of augmented or virtual reality should incorporate head tracking.

VIII. Conclusion

The study evaluated spatial analysis and synthesis methods, specifically SDM, BSDM, and HO-SIRR, along with their variations, in creating synthetic BRIRs of an ITU-R BS.1116-compliant listening room. The objective was to achieve auralization perceptually similar to the KU100 dummy head in terms of spatial fidelity and timbral fidelity, using Zieliński et al.'s fidelity attribute [48]. Certain SDM configurations exhibited spatial fidelity levels comparable to those of HO-SIRR, while BSDM suffered from artifacts in the temporal structure of stimuli. None of the systems were perceptually indistinguishable from the reference. A larger number of microphone capsules enhanced the performance of SDM methods, and a high-quality pressure signal improved timbral and spatial fidelity in certain conditions. The results are expected to more generally describe the performance of synthetic binaural room impulse responses in small rooms, such as critical listening rooms and audio mixing rooms.

Future work can include incorporating head-tracking to evaluate how different rendering methods handle dynamic cues for localization and externalization. Additional areas of study can involve other room types, additional subjective dimensions, and new systems such as 4D-ASDM [22] or REPAIR [23].

Acknowledgment

The authors thank everyone who took part in the listening test and the three anonymous reviewers for their constructive comments, which significantly improved the manuscript's quality and clarity.

References

[1] International Telecommunication Union, *Methods for the Subjective Assessment of Small Impairments in Audio Systems*, Rec. ITU-R BS.1116-3, International Telecommunications Union, Geneva, Switzerland, Feb. 2015.

[2] M. Kleiner, B.-I. Dalenbäck, and P. Svensson, "Auralization-an overview," *J. Audio Eng. Soc.*, vol. 41, no. 11, pp. 861–875, 1993.

[3] E. Pfanzagl-Cardone, "The DOLBY 'Atmos' System," in *The Art and Science of 3D Audio Recording*. Berlin, Germany: Springer, 2023, pp. 143–188.

[4] International Telecommunication Union, *Advanced Sound System for Programme Production*, Rec. ITU-R BS.2051-3, International Telecommunications Union, Geneva, Switzerland, May 2022.

[5] F. Brinkmann, L. Aspöck, D. Ackermann, S. Lepa, M. Vorländer, and S. Weinzierl, "A round robin on room acoustical simulation and auralization," *J. Acoustical Soc. Amer.*, vol. 145, no. 4, pp. 2746–2760, Apr. 2019.

[6] L. Müller and J. Ahrens, "Perceptual differences for modifications of the elevation of early room reflections," in *Proc. Audio Eng. Soc. Conf.: AES 2022 Int. Audio Virtual Augmented Reality Conf.*, Aug. 2022, pp. 1–10.

[7] F. Bederna, L. Müller, and J. Ahrens, "Perceptual detection thresholds for alterations of the azimuth of early room reflections," in *Proc. Jahrestagung Für Akustik*, DAGA 2023, Hamburg, Germany, Mar. 2023, pp. 1–4.

[8] 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.

[9] 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.

[10] S. Tervo, "SDM toolbox," Sep. 2023. [Online]. Available: https://uk.mathworks.com/matlabcentral/fileexchange/56663-sdm-toolbox

[11] J. Pätynen, S. Tervo, and T. Lokki, "Analysis of concert hall acoustics via visualizations of time-frequency and spatiotemporal responses," *J. Acoustical Soc. Amer.*, vol. 133, no. 2, pp. 842–857, 2013.

[12] 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.

[13] S. V. Amengual Garí, J. Pätynen, and T. Lokki, "Physical and perceptual comparison of real and focused sound sources in a concert hall," *J. Audio Eng. Soc.*, vol. 64, no. 12, pp. 1014–1025, Dec. 2016.

[14] S. V. Amengual Garí, D. Eddy, M. Kob, and T. Lokki, "Real-time auralization of room acoustics for the study of live music performance," in *Proc. Jahrestagung Für Akustik*, 2016, pp. 1–4.

[15] O. C. Gomes, N. Meyer-Kahlen, W. Lachenmayr, and T. Lokki, "Perceptual consequences of direction and level of early reflections in a chamber music hall," in *Proc. Jahrestagung Für Akustik*, 2022, pp. 1–4.

[16] P. Coleman, A. Franck, P. Jackson, R. J. 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.

[17] 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, Jun. 2020.

[18] M. Zaunschirm, M. Frank, and F. Zotter, "BRIR synthesis using first-order microphone arrays," in *Proc. Audio Eng. Soc. Conv. 144*, 2018, pp. 1–10.

[19] M. Zaunschirm, M. Frank, and F. Zotter, "Binaural rendering with measured room responses: First-order ambisonic microphone vs. dummy head," *Appl. Sci.*, vol. 10, no. 5, 2020, Art. no. 1631.

[20] S. V. Amengual Garí, J. M. Arend, P. T. Calamia, and P. W. Robinson, "Optimizations of the spatial decomposition method for binaural reproduction," *J. Audio Eng. Soc.*, vol. 68, no. 12, pp. 959–976, 2021.

[21] C. Hold, L. McCormack, and V. Pulkki, "Parametric binaural reproduction of higher-order spatial impulse responses," in *Proc. 24th Int. Congr. Acoust.*, Oct. 2022, pp. 1–5.

[22] E. Hoffbauer and M. Frank, "Four-directional ambisonic spatial decomposition method with reduced temporal artifacts," *J. Audio Eng. Soc.*, vol. 70, no. 12, pp. 1002–1014, Dec. 2022.

[23] L. McCormack, N. Meyer-Kahlen, and A. Politis, "Spatial reconstruction-based rendering of microphone array room impulse responses," *J. Audio Eng. Soc.*, vol. 71, no. 5, pp. 267–280, 2023.

[24] 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.

[25] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoustical Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.

[26] J. Ahrens, "Perceptual evaluation of binaural auralization of data obtained from the spatial decomposition method," in *2019 IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2019, pp. 65–69.

[27] S. V. Amengual Garí, W. Lachenmayr, and E. Mommertz, "Spatial analysis and auralization of room acoustics using a tetrahedral microphone," *J. Acoustical Soc. Amer.*, vol. 141, no. 4, pp. EL369–EL374, 2017.

[28] A. Pawlak, H. Lee, A. Mäkivirta, and T. Lund, "Subjective evaluation of spatial analysis and synthesis methods using different microphone arrays," in *2021 Immersive 3D Audio: From Architecture Automotive*, 2021, pp. 1–7.

[29] D. P. Jarrett, E. A. Habets, and P. A. Naylor, *Theory and Applications of Spherical Microphone Array Processing, Ser. Springer Topics in Signal Processing*, vol. 9. Cham, Switzerland: Springer, 2017.

[30] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing, Ser. Springer Topics in Signal Processing*, vol. 1. Berlin, Germany: Springer, 2008.

[31] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[32] L. Zhang and X. Wu, "On cross correlation based-discrete time delay estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2005, vol. 4, pp. iv/981–iv/984.

[33] J. Yli-Hietanen, K. Kalliojarvi, and J. Astola, "Low-complexity angle of arrival estimation of wideband signals using small arrays," in *Proc. IEEE 8th Workshop Stat. Signal Array Process.*, 1996, pp. 109–112.

[34] Y. Yamasaki and T. Itow, "Measurement of spatial information in sound fields by closely located four point microphone method," *J. Acoustical Soc. Jpn.*, vol. 10, no. 2, pp. 101–110, 1989.

[35] A. Abdou and R. W. Guy, "Spatial information of sound fields for room-acoustics evaluation and diagnosis," *J. Acoustical Soc. Amer.*, vol. 100, no. 5, pp. 3215–3226, Nov. 1996.

[36] F. J. Fahy, *Sound Intensity*, 2nd ed. London, U.K.: CRC Press, 1995.

[37] D. P. Jarrett, E. A. Habets, and P. A. Naylor, "3D source localization in the spherical harmonic domain using a pseudointensity vector," in *2010 IEEE 18th Eur. Signal Process. Conf.*, 2010, pp. 442–446.

[38] J. Ahrens, "Auralization of omnidirectional room impulse responses based on the spatial decomposition method and synthetic spatial data," in *2019 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 146–150.

[39] A. Bassuet, "New acoustical parameters and visualization techniques to analyze the spatial distribution of sound in music spaces," *Building Acoust.*, vol. 18, no. 3/4, pp. 329–347, 2011.

[40] L. McCormack et al., "Applications of spatially localized active-intensity vectors for sound-field visualization," *J. Audio Eng. Soc.*, vol. 67, no. 11, pp. 840–854, Nov. 2019.

[41] L. Gölles and F. Zotter, "Directional enhancement of first-order ambisonic room impulse responses by the 2 2 directional signal estimator," in *Proc. 15th Int. Conf. Audio Mostly*, Sep. 2020, pp. 38–45.

[42] J. Pätynen, S. Tervo, and T. Lokki, "Amplitude panning decreases spectral brightness with concert hall auralizations," in *Proc. Audio Eng. Soc. Conf.: 55th Int. Conf.: Spatial Audio*, 2014, pp. 1–8.

[43] O. Puomio, J. Pätynen, and T. Lokki, "Optimization of virtual loudspeakers for spatial room acoustics reproduction with headphones," *Appl. Sci.*, vol. 7, no. 12, 2017, Art. no. 1282.

[44] S. V. Amengual Garí, "BinauralSDM," Jun. 2023. [Online]. Available: https://github.com/facebookresearch/BinauralSDM

[45] S. Tervo, P. Laukkanen, J. Pätynen, and T. Lokki, "Preferences of critical listening environments among sound engineers," *J. Audio Eng. Soc.*, vol. 62, no. 5, pp. 300–314, 2014.

[46] S. V. Amengual Garí, W. O. Brimijoin, H. G. Hassager, and P. W. Robinson, "Flexible binaural resynthesis of room impulse responses for augmented reality research," in *Proc. EAA Spatial Audio Signal Process. Symp.*, 2019, pp. 161–166.

[47] J. Ahonen and V. Pulkki, "Diffuseness estimation using temporal variation of intensity vectors," in *2009 IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2009, pp. 285–288.

[48] S. K. Zielinski, F. Rumsey, R. Kassier, and S. Bech, "Comparison of basic audio quality and timbral and spatial fidelity changes caused by limitation of bandwidth and by down-mix algorithms in 5.1 surround audio systems," *J. Audio Eng. Soc.*, vol. 53, no. 3, pp. 174–192, 2005.

[49] International Telecommunication Union, *Method for the subjective assessment of intermediate quality level of audio systems*, Rec. ITU-R BS.1534-3, International Telecommunications Union, Geneva, Switzerland, 2014.

[50] K. Hiyama, S. Komiyama, and K. Hamasaki, "The minimum number of loudspeakers and its arrangement for reproducing the spatial impression of diffuse sound field," in *Proc. Audio Eng. Soc. Conv. 113*, 2002, pp. 1–12.

[51] F. Menzer and C. Faller, "Investigations on modeling BRIR tails with filtered and coherence-matched noise," in *Proc. Audio Eng. Soc. Conv. 127*, 2009, pp. 1–9.

[52] D. Johnson and H. Lee, "Huddersfield universal listening test interface generator (HULTI-GEN) version 2," in *Proc. Audio Eng. Soc. Conv. 149*, 2020, pp. 1–4.

[53] A. Farina, "Advancements in impulse response measurements by sine sweeps," in *Proc. Audio Eng. Soc. Conv. 122*, 2007, pp. 1–21, Art. no. 7121.

[54] H. Helmholz, I. Ananthabhotla, P. T. Calamia, and S. V. Amengual Garí, "Towards the prediction of perceived room acoustical similarity," in *Proc. Audio Eng. Soc. Conf.: AES 2022 Int. Audio Virtual Augmented Reality Conf.*, Aug. 2022, pp. 1–11.

[55] T. Surdu, C. Schneiderwind, P. Popp, L. Treybig, and S. Werner, "A. LI. EN: An audiovisual dataset of different acoustical impulse responses measured in a living room environment," in *Proc. Jahrestagung Für Akustik*, DAGA 2023, Hamburg, Germany, 2023, pp. 1644–1647.

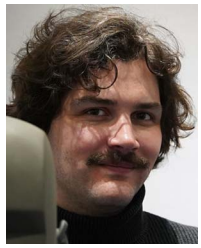
[56] N. Meyer-Kahlen, S. V. Amengual Garí, I. Ananthabhotla, and P. Calamia, "A two-dimensional threshold test for reverberation time and direct-to-reverberant ratio," in *2023 IEEE Immersive 3D Audio: Architecture Automot.*, 2023, pp. 1–8.

[57] L. McCormack, "HO-SIRR," Aug. 2023. [Online]. Available: https://github.com/leomccormack/HO-SIRR

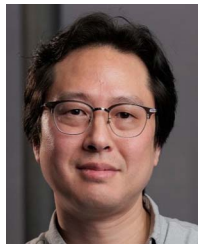
[58] mh acoustics, LLC, "Specification for Eigenbeams" mh acoustics, LLC, Tech. Rep. Version 1 Rev. A., 2016.

[59] B. Bernschütz, "A spherical far field HRIR/HRTF compilation of the Neumann KU 100," in *Proc. 40th Italian Annu. Conf. Acoust. 39th German Annu. Conf. Acoust. Conf. Acoust.*, 2013, pp. 592–595.

[60] P. Majdak, F. Zotter, F. Brinkmann, J. De Muynke, M. Mihocic, and M. Noisternig, "Spatially oriented format for acoustics 2.1: Introduction and recent advances," *J. Audio Eng. Soc.*, vol. 70, no. 7/8, pp. 565–584, 2022.
[61] Bang & Olufsen, "Music for archimedes," [Audio CD], Denmark, 1992.
[62] V. Hansen and G. Munch, "Making recordings for simulation tests in the Archimedes project," *J. Audio Eng. Soc.*, vol. 39, no. 10, pp. 768–774, 1991.
[63] Various Artists, "Anechoic orchestral music recording," [Audio CD] Denon, PG-6006, Nippon Columbia Co., Ltd, Japan, 1988.
[64] H. Lee, M. Frank, and F. Zotter, "Spatial and timbral fidelities of binaural ambisonics decoders for main microphone array recordings," in *Proc. Audio Eng. Soc. Conf.: 2019 AES Int. Conf. Immersive Interactive Audio*, 2019, pp. 1–8.
[65] O. Kirkeby and P. A. Nelson, "Digital filter design for inversion problems in sound reproduction," *J. Audio Eng. Soc.*, vol. 47, no. 7/8, pp. 583–595, 1999.
[66] Z. Schärer and A. Lindau, "Evaluation of equalization methods for binaural signals," in *Audio Engineering Society Convention 126*. Audio Engineering Society, 2009.
[67] S. Holm, "A simple sequentially rejective multiple test procedure," *Scand. J. Statist.*, vol. 6, pp. 65–70, 1979.
[68] A. Field, J. Miles, and Z. Field, *Discovering Statistics Using R*. New York, NY, USA: Sage pub., 2012.
[69] J. Hebrank and D. Wright, "Spectral cues used in the localization of sound sources on the median plane," *J. Acoustical Soc. Amer.*, vol. 56, no. 6, pp. 1829–1834, 1974.
[70] F. Asano, Y. Suzuki, and T. Sone, "Role of spectral cues in median plane localization," *J. Acoustical Soc. Amer.*, vol. 88, no. 1, pp. 159–168, 1990.
[71] E. H. Langendijk and A. W. Bronkhorst, "Contribution of spectral cues to human sound localization," *J. Acoustical Soc. Amer.*, vol. 112, no. 4, pp. 1583–1596, 2002.
[72] P. Zahorik, "Direct-to-reverberant energy ratio sensitivity," *J. Acoustical Soc. Amer.*, vol. 112, no. 5, pp. 2110–2117, Nov. 2002.
[73] H. Lee and D. Johnson, "3D microphone array comparison: Objective measurements," *J. Audio Eng. Soc.*, vol. 69, no. 11, pp. 871–887, 2021.
[74] International Organization for Standardization, *Acoustics – Measurement of Room Acoustic Parameters – Part 1: Performance Spaces*, ISO Standard 3382-1:2009, 2009.
[75] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*. Cambridge, MA, USA: MIT Press, 1997.
[76] A. Andreopoulou and B. F. G. Katz, "Identification of perceptually relevant methods of inter-aural time difference estimation," *J. Acoustical Soc. Amer.*, vol. 142, no. 2, pp. 588–598, Aug. 2017.
[77] T. Hidaka, L. L. Beranek, and T. Okano, "Interaural cross-correlation, lateral fraction, and low- and high-frequency sound levels as measures of acoustical quality in concert halls," *J. Acoustical Soc. Amer.*, vol. 98, no. 2, pp. 988–1007, Aug. 1995.
[78] Z. Ben-Hur, D. L. Alon, R. Mehra, and B. Rafaely, "Binaural reproduction based on bilateral ambisonics and ear-aligned HRTFs," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 901–913, 2021.
[79] A. Gabrielsson, B. Hagerman, T. Bech-Kristensen, and G. Lundberg, "Perceived sound quality of reproductions with different frequency responses and sound levels," *J. Acoustical Soc. Amer.*, vol. 88, no. 3, pp. 1359–1366, 1990.
[80] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*, vol. 22. Berlin, Germany: Springer, 2013.
[81] J. Nikunen and T. Virtanen, "Time-difference of arrival model for spherical microphone arrays and application to direction of arrival estimation," in *2017 IEEE 25th Eur. Signal Process. Conf.*, 2017, pp. 1255–1259.
[82] T. Ziemer, "Source width in music production. Methods in stereo, ambisonics, and wave field synthesis," in *Stud. in Musical acoust. Psychoacoustics*. Berlin, Germany: Springer, 2017, pp. 299–340.
[83] B. Stahl and S. Riedel, "Perceptual comparison of dynamic binaural reproduction methods for sparse head-mounted microphone arrays," *J. Audio Eng. Soc.. Audio Eng. Soc.*, vol. 73, no. 7/8, pp. 442–454, Jul. 2024.
[84] D. R. Begault, E. M. Wenzel, and M. R. Anderson, "Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source," *J. Audio Eng. Soc.*, vol. 49, no. 10, pp. 904–916, 2001.



Alan Pawlak received the degree (with first-class Hons.) in music technology from the University of Huddersfield, Huddersfield, U.K., where he is currently working toward the Ph.D. degree with the Applied Psychoacoustics Laboratory (APL), specializing in spatial audio and binaural rendering. He was an intern with Fraunhofer IIS, Germany, and ZYLIA, Poland, contributing to 3D audio and Higher-Order Ambisonics research during his studies. His Master's research at APL involved developing audio plugins using C++ and JUCE. His research focuses on enhancing binaural auralization quality using spatial room impulse responses. He is currently collaborating on a joint venture Ph.D. degree project with Genelec Oy, Finland. He was the recipient of the awards for Best Music Technology Student and Best Final Year Project with the University of Huddersfield.



Hyunkook Lee received the bachelor's degree in music and sound recording (Tonmeister) and the Ph.D. degree in spatial audio psychoacoustics from the University of Surrey, Guildford, U.K. He is currently the Professor of audio and psychoacoustic engineering and also the Director with Applied Psychoacoustics Laboratory (APL), University of Huddersfield, Huddersfield, U.K. His research interests include perception, capture, and reproduction of spatial audio, and focuses on six-degrees-of-freedom audio perception, binaural audio rendering, and the modeling of immersive auditory experiences. He was a Senior Research Engineer with LG Electronics, working on MPEG audio codec standardization and spatial audio for mobile devices.

Aki Mäkivirta received the Master of Science, Licentiate of Science, and Doctor of Science in Technology degrees in electrical engineering from the Tampere University of Technology, Tampere, Finland, in 1985, 1989, and 1992, respectively. He is currently the R&D Director with Genelec Oy, Iisalmi, Finland. He is also a Fellow of the Audio Engineering Society and a Life Member of the Acoustical Society of Finland.

Thomas Lund has authored papers on human perception, spatialization, loudness, sound exposure and true-peak level. He is currently a Researcher with Genelec Oy, Iisalmi, Finland, and also a Convenor of Working Group TC108X/WG03 under the European Commission. Out of a medical background, he was in Healthcare and as CTO with TC Electronic.