OPEN ACCESS Freely available online

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 May).
DOI: https://doi.org/10.17743/jaes.2022.0072

Spatial Reconstruction-Based Rendering of Microphone Array Room Impulse Responses

LEO MCCORMACK,[1] *AES Student Member*, **NILS MEYER-KAHLEN,**[1] *AES Student Member* **AND**
(leo.mccormack@aalto.fi) (nils.meyer-kahlen@aalto.fi)

ARCHONTIS POLITIS,[2] *AES Member*
(archontis.politis@tuni.fi)

[1] *Department of Information and Communications Engineering, Aalto University, Espoo, Finland*
[2] *Faculty of Information Technology and Communication Sciences, Tampere University, Finland*

A reconstruction-based rendering approach is explored for the task of imposing the spatial characteristics of a measured space onto a monophonic signal while also reproducing it over a target playback setup. The foundation of this study is a parametric rendering framework, which can operate either on arbitrary microphone array room impulse responses (RIRs) or Ambisonic RIRs. Spatial filtering techniques are used to decompose the input RIR into individual reflections and anisotropic diffuse reverberation, which are reproduced using dedicated rendering strategies. The proposed approach operates by considering several hypotheses involving different rendering configurations and thereafter determining which hypothesis reconstructs the input RIR most faithfully. With regard to the present study, these hypotheses involved considering different potential reflection numbers. Once the optimal number of reflections to render has been determined over time and frequency, the array directional responses used to reconstruct the input RIR are substituted with spatialization gains for the target playback setup. The results of formal listening experiments suggest that the proposed approach produces renderings that are perceptually more similar to reference responses, when compared with the use of an established subspace-based detection algorithm. The proposed approach also demonstrates similar or better performance than that achieved with existing state-of-the-art methods.

0 INTRODUCTION

The parameterization and reproduction of microphone array room impulse responses (RIRs) has found application within a number of different areas, including in the technical [1] and perceptual [2] acoustical analysis of concert halls and historical buildings [3] and in artistic productions [4]. More recently, microphone array RIR processing has also been studied in the context of emerging interactive virtual and augmented reality applications [5–7]. Microphone array RIRs may be captured using a spherical microphone array (SMA) or, indeed, any arrangement of microphones over an arbitrary geometry in the general case. The most commonly estimated spatial parameter in existing spatial RIR rendering methods [8, 9] is the direction-of-arrival (DoA) of the direct sound and reflections over time. This information may be used for visualizing and analyzing reflection paths in a room or for auralizing the spatial RIR over a target playback system.

The primary focus of this article pertains to this latter auralization task, whereby a rendering method is first used to convert the input array RIR into a new multichannel RIR, which corresponds to the target playback system. This then allows the acoustical characteristics of the measured space to be imposed onto a monophonic input signal, after it has been convolved with this synthesized RIR and subsequently delivered over the playback system.

0.1 Parametric Spatial RIR Processing

One of the pioneering studies regarding the parameterization of spatial RIRs was conducted by Yamasaki and Itow in [10]. The authors proposed that a microphone array could be used to determine the DoAs of room reflections, through either time difference of arrival (TDoA) or intensity vector (IV) analysis. This information was then used for visualizing the sound propagation paths from the perspective of the receiver position. However, using this information for auralization purposes was not formally investigated until over a decade later.

Methods that aim to auralize a parameterized input spatial RIR can differ in three key respects: 1) most fundamentally, the selection of the underlying sound field model, which describes the assumptions that are made regarding the composition of the microphone array RIR and defines the parameters used to describe it; 2) the input format employed, which can either be the microphone array RIR directly, or an intermediate representation of it (such as Ambisonics [11]); and 3) the algorithmic choices and techniques applied for estimating the model parameters, and for subsequently rendering an RIR for a given target playback setup.

The first spatial RIR processing method used for auralization purposes was the spatial impulse response rendering (SIRR) method [8], which operates based on first-order Ambisonics input and conducts the rendering in a short-time Fourier transform (STFT) domain. The SIRR method assumes that the sound at each time-frequency index originates either from a single direction, from all directions with equal power and random phase (i.e., an isotropic diffuse-field), or from a combination of the two, as determined by an accompanying diffuseness parameter. The method estimates the diffuseness parameter and the DoAs of reflections based on IV analysis. It then relies on mapping directional components to the loudspeakers of a playback system using vector-base amplitude panning (VBAP) [12] and synthesizing isotropic diffuse sounds through the use of decorrelation. More recently, a higher-order formulation of SIRR (HO-SIRR) was proposed in [13], which is based on sector-analysis principles [14, 15] and has a degree of parity with its signal-domain counterpart: higher-order directional audio coding [16, 17]. HO-SIRR aims to estimate the DoA and diffuseness parameters within multiple directionally constrained sectors on the surface of the unit sphere. Therefore, it can resolve simultaneous reflections in the same time-frequency tile, provided that they fall within different sectors. It may also approximate anisotropic energy distributions in the later (more diffuse) part of the response.

The spatial decomposition method (SDM) proposed in [9] is based upon a much simpler sound field model. It assumes the presence of one broadband sound event per time index throughout the response. The method relies on using an open array of omnidirectional microphones and employs DoA estimation based on the TDoA approach, which is conducted within short analysis windows and with a hop size of one sample. As an alternative, broadband IV analysis, which operates on first-order Ambisonics input in the same manner as SIRR, was also explored in [18, 19] and integrated into the SDM MATLAB toolbox [20]. After DoA estimation, SDM performs sample-wise mapping of one of the (omnidirectional) microphone RIRs to the loudspeaker channel responses. SDM is a common choice for technical analysis of impulse responses, with the broadband estimates at high temporal resolution being relatively easy to visualize and interpret. However, although the sound field assumption of one DoA per sample may hold true in the early part of the responses, it is heavily violated in the later part. Here, the estimated directional information captures anisotropy in a more statistical sense [21].

Several other methods, which instead use steered response power (SRP) analysis for estimating the DoAs of reflections within spatial RIRs, have also been explored in [22–24]. Furthermore, a model involving two simultaneous reflections applied to first-order Ambisonic RIRs was shown to provide higher-quality rendering than the single reflection model of SDM in [25]. A largely nonparametric time-domain approach for filtering spatial RIRs to obtain binaural RIRs was also proposed in [26], which used SRP-based DoA estimation to steer the direct sound.

All in all, the relatively simple models forming the basis of SIRR, SDM, and their variants have proved popular in the visualization and auralization of spatial RIRs. However, the more general sector-based model of HO-SIRR, which applies DoA estimation in multiple sectors and can account for anisotropic distributions of diffuse energy, has been shown to provide perceptual benefits compared with methods based on simpler models [13]. In [27], opportunities were also outlined for further extending sound field models for RIR processing, including assuming the presence of several incoming sound events for each time-frequency region, and the use of algorithms such as the multiple signal classification (MUSIC) [28] approach for DoA estimation, which operates based upon the noise subspace of the array spatial covariance matrix (SCM). Subspace-based processing has also recently been explored for decomposing a spatial RIR into directional and residual components, without needing to perform DoA estimation [29].

0.2 Proposed Method

The method explored in the present study is based on formulating a general sound field model in terms of SCMs and is akin to models used in recent parametric rendering techniques applied to continuous signals as input, such as the coding and multi-directional parameterisation of ambisonic sound scenes (COMPASS) method [30]. COMPASS was originally formulated in the spherical harmonic (SH) domain, but it has also been adapted for application in the space-domain more recently in [31]. The model assumes the presence of a variable number of simultaneous sources (or reflections, in the case of RIRs) per time-frequency tile, which are accompanied by an anisotropic diffuse reverberation component.

Spatial analysis and subsequent filtering techniques are then employed to identify and isolate the individual reflections, and a residual component is obtained by subtracting these reflections from the input. These individual components are then reproduced over the target playback setup using dedicated rendering strategies for each.

The spatial reconstruction-based optimization proposed in this article essentially operates by 1) configuring and applying many different spatial analyses and/or rendering techniques in parallel; 2) calculating how well each of these rendering hypotheses can reconstruct the original array RIR; and 3) selecting the optimal rendering configuration per time-frequency tile, prior to synthesizing a RIR corresponding to the target playback setup. In this article, the chosen rendering hypotheses were to consider different numbers of possible reflections. The motivation for selecting this particular aspect of the rendering for optimization, is that an incorrect determination in the reflection number is likely to lead to errors in the DoA estimation. This may subsequently result in poor separation of reflections from the reverberation (modelled by the residual component), which may affect the perceived accuracy and robustness of the rendering. It is also worth bearing in mind that even if the true number of reflections is known (or estimated correctly), it may not be possible to sufficiently isolate them due to limited spatial selectivity of the employed microphone array and beamformer design. This may lead to perceptual issues during the final rendering. Therefore, if auralization is indeed the intended application of such an approach, it may be more appropriate to conduct the rendering using an alternative parameterization in such cases, which the proposed optimization will also seek to find.

This article, therefore, has two main contributions.[1] First, a generalized multidirectional spatial RIR analysis and rendering framework is presented,[2] which operates based on either microphone array or Ambisonic RIRs. The second contribution is the proposed spatial reconstruction-based approach, which is intended to optimize and improve the perceived accuracy of the rendering framework. A formal perceptual experiment comprising three parts is then used to evaluate the performance of the proposed optimization, and to compare the overall rendering method against existing state-of-the-art alternatives.

This article is arranged as follows: SEC. 1 describes the adopted multidirectional sound field model. The proposed analysis framework, which uses a baseline reflection number estimator, is then formulated in SEC. 2. The rendering strategies used to synthesize RIRs corresponding to an arbitrary loudspeaker array are then described in SEC. 3. The proposed rendering optimization, based on the principle of spatial reconstruction, is then described in SEC. 4. The conducted perceptual experiments are then described in SEC. 5, with the results and discussion presented in SEC. 6. The article is concluded in SEC. 7.

1 SOUND FIELD MODEL

It is first assumed that an input Q-channel microphone array RIR has been transformed into the time-frequency domain $\mathbf{x}(t, f) \in \mathbb{C}^{Q\times 1}$, where t and f denote the time and frequency indices, respectively. The SCM of the RIRs are then obtained as $\mathbf{C}_{\mathrm{x}}(t, f) = \mathbb{E}[\mathbf{x}(t, f)\mathbf{x}^{\mathrm{H}}(t, f)] \in \mathbb{C}^{Q\times Q}$, where $\mathbb{E}[.]$ denotes the expectation operator. In practice, expectation is often determined through an averaging operation conducted over a number of temporal frames and sometimes also carried out over frequency groupings, such as octave bands.

It is then assumed that a number $K < Q$ of simultaneous reflections $\mathbf{s}$, at each time-frequency index, are incident from directions $\boldsymbol{\Gamma}_K = [\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_K]$, where $\boldsymbol{\gamma}_k \in S^2$ are Cartesian coordinates of unit length describing the direction of the k^{th} reflection. The input RIR is therefore described as

$$\begin{aligned}\mathbf{x}(t, f) &= \mathbf{x}_{\mathrm{s}}(t, f) + \mathbf{x}_{\mathrm{d}}(t, f) + \mathbf{x}_{\mathrm{n}}(t, f),\\ &= \mathbf{A}_{\mathrm{s}}(f)\mathbf{s}(t, f) + \mathbf{x}_{\mathrm{d}}(t, f) + \mathbf{x}_{\mathrm{n}}(t, f),\end{aligned} \tag{1}$$

where $\mathbf{A}_{\mathrm{s}} = [\mathbf{a}(\boldsymbol{\gamma}_1), ..., \mathbf{a}(\boldsymbol{\gamma}_K)] \in \mathbb{C}^{Q\times K}$ is a matrix of array transfer functions, which are assumed to be known and may be derived from analytical descriptions of the array geometry, numerical simulations, or array calibration measurements; $\mathbf{x}_{\mathrm{s}} \in \mathbb{C}^{Q\times 1}$ describes the array component due to captured reflections; $\mathbf{x}_{\mathrm{d}} \in \mathbb{C}^{Q\times 1}$ describes the capture of diffuse sounds; and $\mathbf{x}_{\mathrm{n}} \in \mathbb{C}^{Q\times 1}$ denotes sensor noise.

Note that the diffuse vector $\mathbf{x}_{\mathrm{d}}$ comprises sounds that are spatially uncorrelated but may not necessarily conform to an isotropic energy distribution. These diffuse sounds are modeled using a dense grid $V \gg Q$ of plane-waves $\mathbf{z} \in \mathbb{C}^{V\times 1}$ incident from directions $\boldsymbol{\Gamma}_V = [\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_V]$, which are uniformly distributed over the sphere, as

$$\mathbf{x}_{\mathrm{d}}(t, f) = \mathbf{A}_{\mathrm{z}}(f)\mathbf{z}(t, f), \tag{2}$$

where $\mathbf{A}_{\mathrm{z}} = [\mathbf{a}(\boldsymbol{\gamma}_1), ..., \mathbf{a}(\boldsymbol{\gamma}_V)] \in \mathbb{C}^{Q\times V}$ are the array transfer functions corresponding to the plane-wave directions. These diffuse sound components are mainly expected to model the late part of the RIR, which would likely be poorly characterized by only $K < Q$ plane-waves. However, they should also model scattered sound energy occurring elsewhere in the response. Note that the time-frequency indices are omitted henceforth for brevity of notation, unless required for clarity.

When the array is presented with only distinct reflections, the SCM for the input array RIR is given as

$$\mathbf{C}_{\mathrm{x,s}} = \mathbf{A}_{\mathrm{s}}\mathbf{C}_{\mathrm{s}}\mathbf{A}_{\mathrm{s}}^{\mathrm{H}}, \tag{3}$$

where $\mathbf{C}_{\mathrm{s}} = \mathbb{E}[\mathbf{s}\mathbf{s}^{\mathrm{H}}] \in \mathbb{C}^{K\times K}$ is the SCM for the reflections.

The SCM of the uncorrelated plane-waves modeling diffuse reverberation is given as $\mathbf{C}_{\mathrm{z}} = \mathbb{E}[\mathbf{z}\mathbf{z}^{\mathrm{H}}] \in \mathbb{C}^{V\times V}$. The SCM has a total power of $P_{\mathrm{z}} = \mathbf{tr}[\mathbf{C}_{\mathrm{z}}]$ and a direction-dependent energy distribution, which is described by the diagonal entries of $\mathbf{C}_{\mathrm{z}}$. However, sometimes it is more useful to impose the more restrictive assumption of an isotropic diffuse field. In this case, $\mathbf{C}_{\mathrm{z}} = P_{\mathrm{z}}\mathbf{I}_V$ and the array SCM, when capturing this isotropic diffuse field, becomes

$$\mathbf{C}_{\mathrm{x,d}} = \mathbb{E}[\mathbf{x}_{\mathrm{d}}\mathbf{x}_{\mathrm{d}}^{\mathrm{H}}] = \mathbf{A}_{\mathrm{z}}\mathbf{C}_{\mathrm{z}}\mathbf{A}_{\mathrm{z}}^{\mathrm{H}} = P_{\mathrm{z}}\mathbf{D}, \tag{4}$$

where $\mathbf{D} = \mathbf{A}_{\mathrm{z}}\mathbf{A}_{\mathrm{z}}^{\mathrm{H}} \in \mathbb{C}^{Q\times Q}$ is the diffuse coherence matrix (DCM) for the employed microphone array, which describes the relative inter-channel relationships incurred when the array configuration captures a diffuse-field.

The array SCM when capturing only sensor noise, which is assumed to be uncorrelated and of equal power P_{n} across all sensors, is expressed as

$$\mathbf{C}_{\mathrm{x,n}} = \mathbb{E}[\mathbf{x}_{\mathrm{n}}\mathbf{x}_{\mathrm{n}}^{\mathrm{H}}] = P_{\mathrm{n}}\mathbf{I}_Q. \tag{5}$$

The total array SCM is therefore

$$\mathbf{C}_{\mathrm{x}} = \mathbf{C}_{\mathrm{x,s}} + \mathbf{C}_{\mathrm{x,d}} + \mathbf{C}_{\mathrm{x,n}}. \tag{6}$$

[1] A MATLAB implementation of the parametric rendering framework, including the proposed reconstruction-based optimization for reflection number estimation, may be found here: https://github.com/leomccormack/REPAIR

[2] Note that a preliminary version of the employed parametric rendering framework was presented in [32].

1.1 Spherical Harmonic Domain

The aforementioned sound field model is also directly applicable in the SH domain [33]. In this case, the input RIRs are Ambisonic RIRs of order N, with $Q = (N + 1)^2$ channels. If an ideal SH receiver is employed, then the matrices of array transfer functions $\mathbf{A}_\mathrm{s}$ and $\mathbf{A}_\mathrm{z}$ may be substituted with matrices of broad-band real-valued SH weights $\mathbf{Y}_\mathrm{s} \in \mathbb{R}^{(N+1)^2 \times K}$ and $\mathbf{Y}_\mathrm{z} \in \mathbb{R}^{(N+1)^2 \times V}$, respectively. Note that in this ideal SH case, because of the orthonormality of the SHs, the DCM in Eq. (4) becomes an identity matrix.

However, since ideal SH receivers are only obtainable through simulations, Ambisonic RIRs of real spaces are instead obtained through microphone array measurements. Here, a frequency-dependent encoding matrix [34, 35, 33] $\mathbf{E} \in \mathbb{C}^{(N+1)^2 \times Q}$, is used to convert a microphone array RIR into the SH domain as $\tilde{\mathbf{x}} = \mathbf{E}\mathbf{x}$. The resulting SH components will, however, succumb to spatial aliasing above a certain frequency limit and may also be corrupted by sensor noise (especially at low frequencies and higher orders). The frequency bandwidths at which usable SH components may be obtained (spatially speaking) is order-dependent for SMAs with uniform sensor placement [36] and also direction-dependent in the case of nonuniform and/or nonspherical arrays. Therefore, the broadband $\mathbf{Y}_\mathrm{s}$ and $\mathbf{Y}_\mathrm{z}$ vectors may be more suitably replaced with $\mathbf{E}\mathbf{A}_\mathrm{s}$ and $\mathbf{E}\mathbf{A}_\mathrm{z}$, in order to better model these behaviors [37]. Note that although the DCM in this case preserves a diagonal structure up to the spatial aliasing limit of the array, the noise SCM becomes nondiagonal due to the encoding process $\tilde{\mathbf{C}}_{\mathrm{x,n}} = P_\mathrm{n}\mathbf{E}\mathbf{E}^\mathrm{H}$ [38].

In Sec. 5, the proposed method is evaluated when using both microphone array RIRs and Ambisonic RIRs as input. For these latter test cases, the encoding filter matrix $\mathbf{E}$ is computed following the design described in [39]. First, this involves expanding the array transfer functions into SH coefficients. This is then followed by applying a regularized least-squares solution to map the array response coefficients to the SH coefficients with

$$\mathbf{S}_\mathrm{z} = \mathbf{A}_\mathrm{z}\mathbf{Y}_\mathrm{z}^\mathrm{T}\left[\mathbf{Y}_\mathrm{z}\mathbf{Y}_\mathrm{z}^\mathrm{T}\right]^{-1}, \tag{7}$$

$$\mathbf{E} = \hat{\mathbf{S}}_\mathrm{z}^\mathrm{H}\left[\mathbf{S}_\mathrm{z}\mathbf{S}_\mathrm{z}^\mathrm{H} + \beta^2\mathbf{I}_Q\right]^{-1}, \tag{8}$$

where $\mathbf{Y}_\mathrm{z}$ is computed up to the maximum order of the employed measurement grid (rather than the maximum order of the array); $\hat{\mathbf{S}}_\mathrm{z} \in \mathbb{C}^{Q \times (N+1)^2}$ denotes a truncated version of $\mathbf{S}_\mathrm{z}$, in order to retain only the first $(N + 1)^2$ columns; and $\beta > 0$ is a regularization parameter.

2 SPATIAL ANALYSIS

This section describes the spatial parameter analysis techniques used for the present study. SCM preprocessing operations are first described, which allow certain SH domain methods [30] to be generalized and applied directly to arbitrary microphone array input. A baseline reflection number estimator, which will be compared with the proposed spatial reconstruction–based approach later in this article, is also presented.

2.1 SCM Frequency Averaging and Coherent Focusing

In the case of parametric methods operating on running signals, the intended applications typically involve the analysis of sound scenes comprising multiple sound sources, which are assumed to be uncorrelated. Therefore, higher-frequency resolution is typically favored for such methods in practice, with the assumption that the sound sources are likely to be sparse across the time-frequency representation. In the case of RIR analysis, however, the sound sources of interest are room reflections corresponding to a single source–receiver combination, with their density rapidly increasing with time. Therefore, configuring the selected time-frequency transform to adopt a lower frequency resolution, but with a higher temporal resolution, may be more beneficial for the present application.

It is also highlighted that early reflections may be largely viewed as being replicas of the direct sound, except with different temporal offsets and magnitudes. As time progresses, it becomes increasingly likely that these partially coherent signals will fall within the same analysis window, which reduces the effective rank of the SCMs and can degrade the performance of subspace-based reflection number and DoA estimation methods. Therefore, certain techniques have been proposed to address this issue by restoring the effective rank of the SCMs in the presence of coherent sources; these include frequency [40], temporal [41], and spatial smoothing [42].

In this work, it is assumed that an STFT is employed, which provides a uniform frequency sampling. The WINGS coherent focusing method [40, 43] is then applied, in order to group and average the array SCMs into octave bands. These frequency-averaged array SCMs, which are aligned to each octave-band center frequency f_0, are obtained as

$$\mathbf{C}_\mathrm{x}^{(\mathrm{OCT})}(f_0) = \sum_{f_i = f_l}^{f_u} \mathbf{T}_\mathrm{coh}(f_i, f_0)\mathbf{C}_\mathrm{x}(f_i)\mathbf{T}_\mathrm{coh}^\mathrm{H}(f_i, f_0), \tag{9}$$

where f_l and f_u denote the lower and upper frequency indices, respectively, which define the octave-band grouping, and $\mathbf{T}_\mathrm{coh} \in \mathbb{C}^{Q \times Q}$ is the coherent focusing matrix computed as [40, 43]

$$\mathbf{T}_\mathrm{coh}(f, f_0) = [\mathbf{A}_\mathrm{z}(f_0)\mathbf{Y}_\mathrm{z}^\mathrm{T}][\mathbf{A}_\mathrm{z}(f)\mathbf{Y}_\mathrm{z}^\mathrm{T}]^\dagger, \tag{10}$$

where † denotes the Moore–Penrose pseudoinverse. Note that this coherent focusing is only required if the steering vectors are different across frequency. For frequency-independent steering vectors, such as those used for ideal SH responses, simply averaging the array SCMs across frequency (without focusing) should be sufficient to decorrelate the coherent source signals. However, it is noted that if this focusing operation were to be applied to such responses, it would in any case be intrinsically bypassed.

2.2 SCM Whitening

The employed baseline reflection number and DoA estimators are built upon the subspace principles of array signal processing. Typically, such methods rely on analyzing the eigenvalues and/or eigenvectors of the SCMs and adopt a

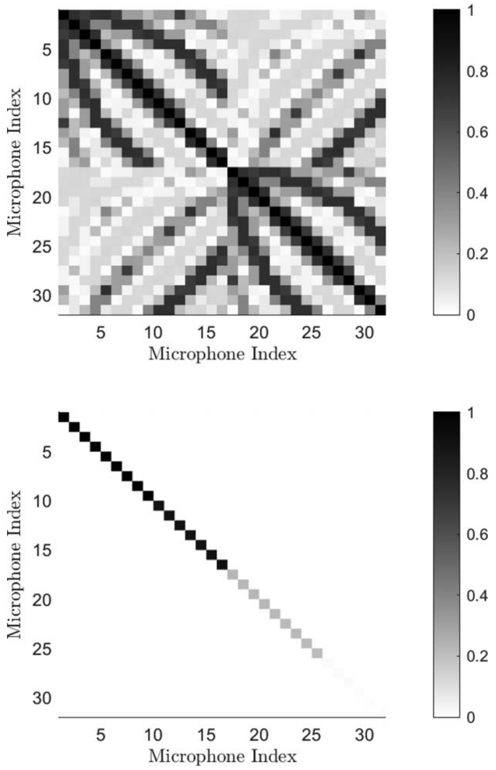

Fig. 1. Normalized DCM for the eigenmike m32 in the octave band around $f = 2$ kHz (a) before and (b) after spatial whitening.

simpler sound field model than the one described in SEC. 1. These methods often assume that directional signals are accompanied only by uncorrelated sensor noise. The methods therefore rely on the array SCMs conforming to an identity-like structure, with its eigenvalues approximately all being equal to P_{n}, when no directional signals are active. In the present case, however, reflections are to be detected while also in the presence of diffuse reverberation, which does not necessarily produce a diagonal SCM for any arbitrary microphone array configuration. For this reason, and assuming $P_{\mathrm{z}} \gg P_{\mathrm{n}}$, it may be beneficial to first spatially whiten the array SCMs, such that they exhibit this identity-like structure when they capture diffuse conditions instead, as also conducted recently in [31, 44].

The spatial whitening procedure is conducted by first obtaining frequency-averaged DCMs, in a similar manner as in Eq. (9), and decomposing them as

$$\mathbf{D}^{(\mathrm{OCT})}(f_0) = \sum_{f_i=f_l}^{f_u} \mathbf{T}_{\mathrm{coh}}(f_i, f_0)\mathbf{D}(f_i)\mathbf{T}_{\mathrm{coh}}^{\mathrm{H}}(f_i, f_0), \quad (11)$$

$$= \mathbf{R}\mathbf{\Lambda}\mathbf{R}^{\mathrm{H}}, \quad (12)$$

which permits acquiring the spatial whitening matrix as $\mathbf{T}_{\mathrm{w}} = \mathbf{\Lambda}^{-1/2}\mathbf{R}^{\mathrm{H}} \in \mathbb{C}^{Q\times Q}$. Spatially whitened (and frequency-averaged) SCMs may then be obtained with the following

$$\hat{\mathbf{C}}_{\mathrm{x}}^{(\mathrm{OCT})} = \mathbf{T}_{\mathrm{w}}\mathbf{C}_{\mathrm{x}}^{(\mathrm{OCT})}\mathbf{T}_{\mathrm{w}}^{\mathrm{H}}. \quad (13)$$

An example of this operation is depicted in Fig. 1. Here, an SCM for a 32-sensor rigid baffle SMA (radius of 0.042 m), when capturing an isotropic diffuse-field, is shown both with and without the whitening applied, given an octave-band grouping centered around 2 kHz. It is also noted that when using orthonormal basis functions as the array steering vectors (such as broadband SHs), this whitening operation (along with the coherent focusing operation) would also be intrinsically bypassed in such cases.

2.3 Baseline Reflection Number Detection

The detection of the number of reflections may be conducted for each time window and octave-band, by first performing a subspace decomposition of the frequency-averaged and spatially whitened SCMs as

$$\hat{\mathbf{C}}_{\mathrm{x}}^{(\mathrm{OCT})} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{\mathrm{H}} = \sum_{k=1}^{K} \lambda_k \mathbf{v}_k \mathbf{v}_k^{\mathrm{H}} + \sum_{k=K+1}^{Q} \lambda_k \mathbf{v}_k \mathbf{v}_k^{\mathrm{H}}, \quad (14)$$

where $\lambda_1 > ... > \lambda_Q$ are the eigenvalues in descending order, and $\mathbf{v}_k$ are their respective eigenvectors. There are then a variety of different detection algorithms available [45–47], which operate based on these eigenvalues and/or eigenvectors. In this work, the SORTE algorithm [45] was selected as the baseline detection algorithm, which requires that the differences between successive eigenvalues are first calculated as

$$\nabla\lambda_i = \lambda_i - \lambda_{i+1}, \quad \text{for } i = 1, ..., Q-1. \quad (15)$$

Following this, the number of reflections is given by

$$K_{\mathrm{SORTE}} = \arg\min_{k} \ f(k) \quad \text{for } k = 1, ..., Q-3, \quad (16)$$

where

$$f(k) = \begin{cases} \frac{\sigma_{k+1}^2}{\sigma_k^2}, & \sigma_k^2 > 0 \\ +\infty, & \sigma_k^2 = 0 \end{cases}, \quad \text{for } k = 1, ..., Q-2, \quad (17)$$

given the eigenvalue difference variances σ_k^2, which are defined as

$$\sigma_k^2 = \frac{1}{Q-k}\sum_{i=k}^{Q-1}\left(\nabla\lambda_i - \frac{1}{Q-k}\sum_{i=k}^{Q-1}\nabla\lambda_i\right)^2. \quad (18)$$

2.4 Reflection DoA Estimation

With the number of reflections K now detected for each time window and frequency grouping, MUSIC [28] may be employed to estimate their directions. This is realized by first scanning a dense grid of directions as

$$P_{\mathrm{MUSIC}}^{(1)}(\boldsymbol{\gamma}) = \frac{1}{||\mathbf{V}_{\mathrm{n}}^{\mathrm{H}}\mathbf{T}_{\mathrm{w}}\mathbf{a}(\boldsymbol{\gamma}, f_0)||^2}, \quad \text{for } \boldsymbol{\gamma} \in \mathbf{\Gamma}_V, \quad (19)$$

where $\mathbf{V}_{\mathrm{n}} \in \mathbb{C}^{Q\times(Q-K)}$ is the noise subspace, which consists of the eigenvectors corresponding to the lowest $Q - K$ eigenvalues; and $\|.\|$ denotes the Euclidean norm of the enclosed vector. Peaks evident within the resulting spatial pseudospectrum then reveal likely reflection DoAs. The reflection DoA estimates may therefore be subsequently obtained through the application of a peak-finding algorithm applied to this spherical data.

In this work, an iterative process is followed, based on the identification of the maximum peak within the pseudospectrum, determining its direction (expressed as $\hat{\boldsymbol{\gamma}}_k$), and subsequently suppressing this peak through the application of a directional masking function. This procedure was originally used in [27] and later in [30] but is detailed here formally for the first time. The peak suppression of the kth DoA is performed through the application of an inverse Von Mises–Fisher distribution function concentrated around $\hat{\boldsymbol{\gamma}}_k$ as

$$m(\boldsymbol{\gamma}, \hat{\boldsymbol{\gamma}}_k) = \left(\epsilon + \frac{\kappa e^{\kappa \boldsymbol{\gamma}^{\mathrm{T}} \hat{\boldsymbol{\gamma}}_k}}{2\pi e^{\kappa} - e^{-\kappa}}\right)^{-1}, \quad \text{for } \boldsymbol{\gamma} \in \boldsymbol{\Gamma}_V, \tag{20}$$

where ϵ is a small constant, and κ is the concentration parameter. These were empirically set to fixed values of $\epsilon = 10^{-5}$ and $\kappa = 50$ in the present study. The peak finding algorithm at the kth step proceeds as

$$\hat{\boldsymbol{\gamma}}_k = \arg\max_{\boldsymbol{\gamma}} P^{(k)}_{\mathrm{MUSIC}}(\boldsymbol{\gamma}), \tag{21}$$

$$P^{(k+1)}_{\mathrm{MUSIC}}(\boldsymbol{\gamma}) = P^{(k)}_{\mathrm{MUSIC}}(\boldsymbol{\gamma}) m(\boldsymbol{\gamma}, \hat{\boldsymbol{\gamma}}_k), \tag{22}$$

until K reflection DoAs have been extracted.

3 RENDERING FRAMEWORK

This section describes the spatial filtering and spatialization techniques adopted by the employed rendering framework. The framework first uses the detected number of reflections and corresponding DoA estimates to isolate individual reflections. These reflections are then re-encoded and subtracted from the input array response, in order to also obtain an estimate of the anisotropic diffuse components [30]. Therefore, the reflection number estimates will greatly influence the output spatial distribution and balance between the directional and diffuse rendering, thus contributing to the motivation for selecting this particular parameter to optimize in the following section.

3.1 Rendering Reflections

Given an estimated (or postulated) reflection number K, and the corresponding DoAs $\boldsymbol{\Gamma}_K$ for each time-frequency tile, the signals of the direct sound and reflections may be isolated as $\mathbf{s} = \mathbf{W}_{\mathrm{s}}\mathbf{x}$, using a matrix of beamforming weights, $\mathbf{W}_{\mathrm{s}} = [\mathbf{w}(\boldsymbol{\gamma}_1), ..., \mathbf{w}(\boldsymbol{\gamma}_K)] \in \mathbb{C}^{K \times Q}$. In the present study, the following super-directive beamformer design was selected for this task [48]

$$\mathbf{w}(\boldsymbol{\gamma}_k) = \frac{\mathbf{a}^{\mathrm{H}}(\boldsymbol{\gamma}_k)(\mathbf{D} + \beta \mathbf{I}_Q)^{-1}}{\mathbf{a}^{\mathrm{H}}(\boldsymbol{\gamma}_k)(\mathbf{D} + \beta \mathbf{I}_Q)^{-1}\mathbf{a}(\boldsymbol{\gamma}_k)}, \tag{23}$$

where $\beta > 0$ is a regularization parameter to account for cases where the DCM may be singular, which can occur at low frequencies. Note that if the input signals and steering vectors are of broadband SHs, then this design reverts to hyper-cardioid (maximum-directivity) beamforming; since $\mathbf{D} = \mathbf{I}_Q$ in such cases.

The isolated reflections may then be spatialized directly over an L-channel loudspeaker array (with arbitrary loudspeaker directions $\boldsymbol{\Gamma}_L = [\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_L]$) as

$$\mathbf{y}_{\mathrm{s}} = \mathbf{G}_{\mathrm{s}}\mathbf{W}_{\mathrm{s}}\mathbf{x}, \tag{24}$$

where $\mathbf{G}_{\mathrm{s}} = [\mathbf{g}(\boldsymbol{\gamma}_1), ..., \mathbf{g}(\boldsymbol{\gamma}_K)] \in \mathbb{R}^{L \times K}$ is a matrix of VBAP [12] gains, $\mathbf{g}(\boldsymbol{\gamma}_k) = [g_1(\boldsymbol{\gamma}_k), ..., g_L(\boldsymbol{\gamma}_k)]$, which correspond to the same DoAs used to steer the beamformers. Note that because subsequent auralization should be conducted in an anechoic chamber, or binaurally over headphones, the frequency-dependent normalized VBAP gains optimized for free-field conditions, as described in [49], were employed for the present study.

Rather than spatializing the isolated reflections over the target playback setup, one may also replace the VBAP gains with microphone array transfer functions corresponding to the same estimated DoAs. In this case, an estimate of the input microphone array RIR containing only reflections may be obtained as

$$\hat{\mathbf{x}}_{\mathrm{s}} = \mathbf{A}_{\mathrm{s}}\mathbf{W}_{\mathrm{s}}\mathbf{x}. \tag{25}$$

3.2 Rendering Diffuse Reverberation

The isolated direct sound and reflections may then be subtracted from the input array response, in order to obtain the residual component $\hat{\mathbf{x}}_{\mathrm{d}} = \mathbf{W}_{\mathrm{d}}\mathbf{x} = \mathbf{x} - \hat{\mathbf{x}}_{\mathrm{s}}$. This residual component encapsulates any remaining reflections (of typically lower amplitude) and spatially incoherent (diffuse) sounds. The ambient extraction matrix is given as [30, 31]

$$\mathbf{W}_{\mathrm{d}} = \mathbf{I}_Q - \mathbf{A}_{\mathrm{s}}\mathbf{W}_{\mathrm{s}}. \tag{26}$$

Once the ambient array signals have been obtained, they are subsequently reproduced over the target loudspeaker setup with the following

$$\mathbf{y}_{\mathrm{d}} = d_{\mathrm{EQ}}\mathbf{G}_{\mathrm{d}}\hat{\mathbf{x}}_{\mathrm{d}} = d_{\mathrm{EQ}}\mathbf{G}_{\mathrm{d}}\mathbf{W}_{\mathrm{d}}\mathbf{x}, \tag{27}$$

where $d_{\mathrm{EQ}} = \mathbf{tr}[\mathbf{D}/V]^{-1/2}$ is an equalization term used to mitigate possible timbral colorations incurred when capturing diffuse sounds. The matrix $\mathbf{G}_{\mathrm{d}} \in \mathbb{C}^{L \times Q}$ then represents a set of frequency-dependent beamformers, which are oriented towards the loudspeaker directions and are forced to exhibit an energy-preserving property [31]. This is realized by forcing the Hermitian transpose of the stacked array steering vectors (corresponding to the loudspeaker directions $\mathbf{A}_l \in \mathbb{C}^{Q \times L}$) to be unitary with

$$\mathbf{A}_l^{\mathrm{H}} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathrm{H}}, \tag{28}$$

$$\mathbf{G}_{\mathrm{d}} = \frac{1}{\sqrt{V}}\hat{\mathbf{U}}\mathbf{V}^{\mathrm{H}}, \tag{29}$$

where $\hat{\mathbf{U}} \in \mathbb{C}^{L \times Q}$ denotes a truncated version of $\mathbf{U}$ (i.e., retaining only the first Q columns). It is noted that this particular design reverts to the energy-preserving Ambisonic decoder (EPAD) proposed in [50], when using broadband SHs as the array steering vectors.

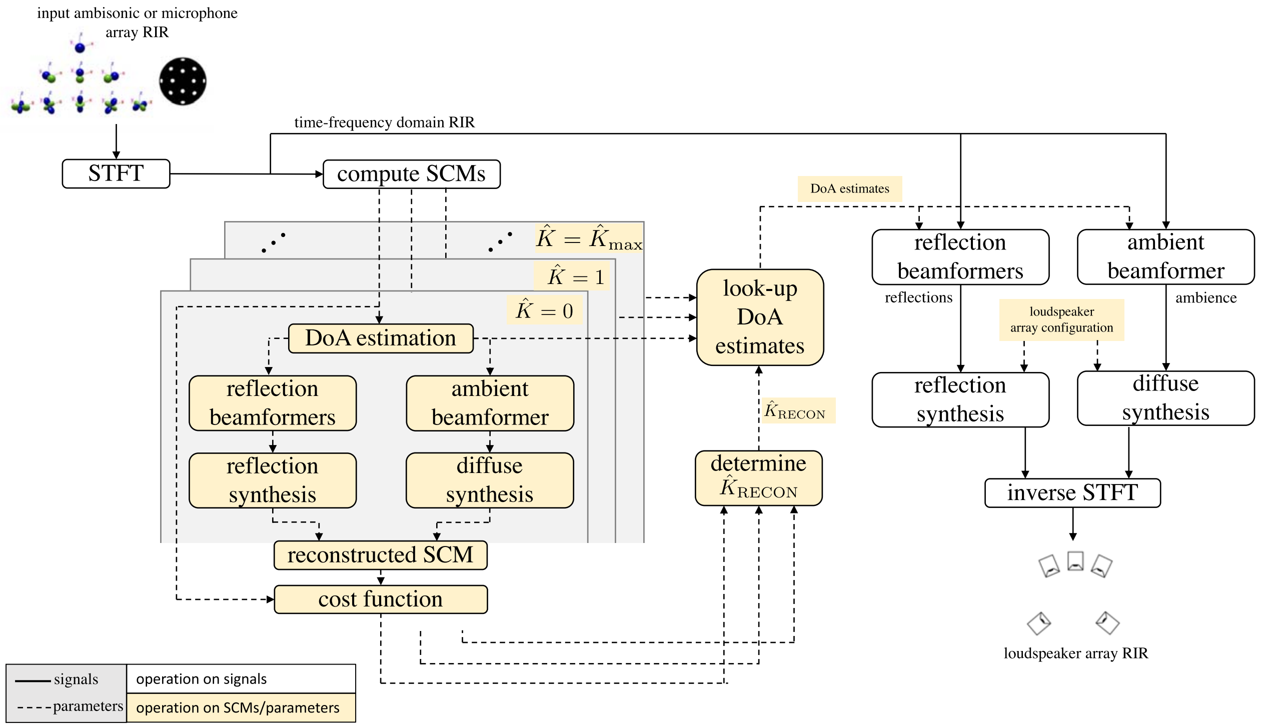

Fig. 2. Block diagram of the proposed spatial reconstruction-based framework. In this case, the framework is configured to test multiple hypotheses regarding the number of reflections. The reflection number that provides the lowest reconstruction error is then chosen when synthesizing the loudspeaker array RIR.

3.3 Overall Rendering and Decorrelation

The output loudspeaker array RIR may be obtained as

$$\mathbf{y}(t, f) = \mathbf{y}_{\text{s}}(t, f) + \mathcal{D}[\mathbf{y}_{\text{d}}(t, f)], \tag{30}$$

where $\mathcal{D}[.]$ denotes a decorrelation operation performed on the enclosed loudspeaker response, in order to enforce diffuse properties. In this work, the decorrelation is realized through simple phase randomization of the RIR time-frequency representation. However, note that further improvements to this diffuse rendering may include a combination of adaptive mixing and minimal use of decorrelators, for example, as described in [51, 52].

4 PROPOSED SPATIAL RECONSTRUCTION-BASED OPTIMIZATION

In this section, the proposed spatial reconstruction-based optimization is described. The optimization is based upon exploring several hypotheses related to the employed spatial analysis and/or rendering techniques and selecting the one that most accurately reconstructs the original array response. Upon determining the optimal rendering configuration for each time-frequency index, the response is then rendered for the reproduction setup.

In principle, these different hypotheses may relate to any aspect(s) of the presented framework. However, in the present study, focus was placed on the reflection number estimation task, since this is the spatial parameter deemed by the authors as most likely to influence the overall performance of the rendering framework. It is also noted that although the baseline SORTE approach, described in SEC. 2.3, is an established method for the detection of active sound sources within continuous signals, it is unclear how well the approach performs for the task of detecting reflections within RIRs. The main areas of concern are that SORTE will always return at least one reflection and may also detect maximal reflection numbers when the array SCM is near full-rank. These traits may lead to suboptimal handling of the late part of the response, where it may be more appropriate to assume that no reflections are active and to only apply diffuse rendering strategies.

Since the cost-function employed by the proposed reconstruction-based optimization operates in the spatial covariance domain, the reconstructed microphone array SCMs are first calculated for each reflection number hypothesis as

$$\begin{aligned}\hat{\mathbf{C}}_{\mathbf{x}}^{(\text{OCT},\hat{K})} &= (\mathbf{A}_{\text{s}}^{(\hat{K})}\mathbf{W}_{\text{s}}^{(\hat{K})})\mathbf{C}_{\text{x}}^{(\text{OCT})}(\mathbf{A}_{\text{s}}^{(\hat{K})}\mathbf{W}_{\text{s}}^{(\hat{K})})^{\text{H}} \\ &\quad + \mathbf{T}_{\text{uw}}\text{Diag}[\mathbf{W}_{\text{d}}^{(\hat{K})}\mathbf{C}_{\text{x}}^{(\text{OCT})}(\mathbf{W}_{\text{d}}^{(\hat{K})})^{\text{H}}]\mathbf{T}_{\text{uw}}^{\text{H}}, \\ &\quad \text{for} \quad \hat{K} \in 0, ..., \lfloor Q/2 \rfloor,\end{aligned} \tag{31}$$

where the superscript $(\hat{K})$ denotes that the rendering matrices were computed based upon the hypothesis that $\hat{K}$ reflections were present; $\mathbf{T}_{\text{uw}} = \mathbf{T}_{\text{w}}^{-1} = \mathbf{R}\mathbf{\Lambda}^{1/2}$ is a un-whitening operation, (the inverse operation of Eq. (13)), which reintroduces the interchannel coherence that the array would naturally capture when under diffuse conditions; and Diag[.] denotes the construction of a diagonal matrix based on the diagonal entries of the enclosed matrix.

The employed cost-function is then simply the squared Frobenius norm, $||.||_F$, of the difference between the original and reconstructed SCMs

$$K_{\text{RECON}} = \underset{\hat{K}}{\arg\min}\, ||\mathbf{C_x}^{(\text{OCT})} - \hat{\mathbf{C}}_{\mathbf{x}}^{(\text{OCT},\hat{K})}||_F^2. \qquad (32)$$

Note that this specific cost-function was selected because it penalizes differences in both the autochannel and the interchannel components of the residual SCM. A block diagram of this proposed reconstruction-based optimization is depicted in Fig. 2.

5 EVALUATION

In order to evaluate the presented rendering framework, when using the spatial reconstruction-based optimization described in Sec. 4, a binaural multiple-stimuli listening experiment was conducted. The experiment was divided into three parts. All three parts of the experiment were based on the use of two simulated shoebox rooms,[3] which employed the use of the image source method [53]. The first room was configured to resemble an acoustically *dry* space, with dimensions 6 × 5 × 3.1 m (Width × Depth × Height) and RT60 times of 0.33, 0.39, 0.26, 0.20, 0.07, and 0.04 s in octave bands of 125 Hz to 4 kHz, whereas the second room was a larger and slightly more reverberant (*rev*) 10 × 6 × 3.5 m space, with RT60 times of 0.52, 0.60, 0.40, 0.20, 0.20, and 0.13 s. The receiver position was set to approximately the center of the room, with the source position placed approximately 2 m in front of it. The motivation for moving the receiver position slightly away from the center of the room and having a small offset in the source position was to mitigate against the perfect simultaneous arrival of image sources. This is because such cases would be unlikely to occur in real scenarios, and certain single-direction approaches, such as first-order SIRR and SDM, may perform suboptimally.

To obtain reference binaural RIRs, all incoming image sources at the receiver position were quantized to a 36-point t-design [54] and then convolved with the respective head-related transfer functions (HRTFs) corresponding to a KU100 dummy-head simulator [55]. These reference binaural RIRs were then convolved with four contrasting stimuli: a kick drum, snare drum, a trombone, and male speech. These particular stimuli were selected as they represent a balance between more transient (kick drum and snare) and more harmonic and stationary (trombone and speech) stimuli. The kick drum, trombone, and speech stimuli were also employed when evaluating HO-SIRR [13], in which the kick drum was found to be especially revealing of potential artefacts incurred during the rendering of spatial RIRs.

The HRTFs were then substituted with microphone array transfer functions, in order to acquire synthetic microphone array RIRs. These were then passed through the different rendering methods under test. Four different receivers were used in this study[4]: a 32-sensor rigid SMA with a radius of 0.042 m, which is the same array configuration used by the commercially available Eigenmike32 (*em32*) [56]; a four-sensor open tetrahedral SMA (*tetra*) with cardioid directivities and a radius 0.02 m, representing a popular array configuration used for the capture of first-order Ambisonics; a uniform six-sensor open SMA with omnidirectional sensors and a radius of 0.025 m, which corresponds to the 3D intensity probe (*ip*) commonly used in conjunction with the SDM method; and an ideal SH receiver of order *N* (*shN*), which does not exhibit any frequency-dependent performance limitations (for example, due to baffle scattering effects and noncoincident sensor placements).

The spatial analysis and rendering framework, as outlined in Secs. 2 and 3, which is henceforth referred to as "*repair*," was implemented using an STFT with a window size of $5.\dot{3}$ ms (256 samples at 48 kHz) and a hop size of $2.\dot{6}$ ms using a Hann window. The SCMs of the input RIRs were averaged over time using a recursive one-pole filter with a coefficient value of 0.5. For the spatial analysis, the SCMs were also averaged over frequency and grouped into octave bands. Note that the maximum number of simultaneous reflections was set to $K_{\max} = \min(\lfloor Q/2 \rfloor, 8)$. The target loudspeaker setup for the rendering was set to the same 36-point t-design, and the output loudspeaker RIRs were subsequently convolved with the reference HRTFs. The resulting binaural RIRs were then convolved with the four monophonic stimuli in order to obtain the audio for each test case. Note that the audio files used for the experiment may also be generated using the open-source toolbox.

The test participants were provided with Sennheiser HD650 headphones and presented with a graphical user interface comprising a number of sliders, one for each method (or rendering configuration) under test. All three tests included a known and hidden reference (*hidden_ref*). The slider scale displayed the verbal anchors *Bad*, *Poor*, *Fair*, *Good*, and *Excellent*, in steps of 20 points from 0 to 100. The test participants were instructed to score the test cases based on their combined spatial and timbral differences with respect to the known reference and to each other and with due consideration given to these verbal anchors. The participants were permitted to loop the presented audio files over shorter time intervals. The same 13 participants took part in all three tests and (excluding breaks between tests) took approximately 50 minutes on average.

5.1 Rendering Configurations and Methods under Test

The purpose of the first part of the listening experiment was to investigate the perceived differences between the use of the baseline reflection number estimator, K_{SORTE} (*sorte*), and the proposed spatial reconstruction-based approach, which was configured to determine the optimal number of reflections to render, K_{RECON} (*recon*). For this test, the *em32* and fourth-order ideal SH receivers were se-

[3] The employed shoebox room simulator may be found here: https://github.com/polarch/shoebox-roomsim

[4] The SMA simulator may be found here: https://github.com/polarch/Array-Response-Simulator

Table 1. Test cases for listening experiment part 1.

Name	Receiver	Algorithm
hidden_ref	Quantized to t-design	Direct binauralization of the quantized image sources
recon_sh4	Fourth-order SH receiver	Presented framework using the spatial reconstruction-based optimization
oracle_sh4	Fourth-order SH receiver	Presented framework using known reflections
sorte_sh4	Fourth-order SH receiver	Presented framework using the SORTE detection algorithm
recon_em32	32-sensor rigid SMA	Presented framework using the spatial reconstruction-based optimization
oracle_em32	32-sensor rigid SMA	Presented framework using known reflections
sorte_em32	32-sensor rigid SMA	Presented framework using the SORTE detection algorithm

lected. For added insight, a ground-truth (*oracle*) test case was included, which bypassed the spatial analysis and applied the rendering techniques based on known reflection data taken from the reference image source echogram. Note that for time hops where the number of reflections in the echogram exceeded $K_{\max}$, the reflections with the highest magnitudes were selected. The test cases for the first part of the experiment are summarized in Table 1.

For the second part of the listening experiment, the proposed method was compared against HO-SIRR, which is an existing multidirectional spatial RIR rendering method. HO-SIRR is formulated in the SH domain, and ideal SH receivers were used in its evaluation in [13], in which HO-SIRR was shown to outperform the majority of other tested rendering methods. The HO-SIRR test cases for the present study were rendered using the open-source HO-SIRR MATLAB toolbox [57] with the default configuration (commit from November 5, 2020). Both *em32* and fourth-order *sh4* receivers were employed for this second test. However, since HO-SIRR does not directly operate in the space-domain, the input RIRs were converted into the SH domain for the em32 case using the encoder matrix described by Eq. (7) (with $\beta = 0.0889$). The test cases for the second part of the experiment are summarized in Table 2.

Note that the first two parts of the experiment focused on the use of the em32 microphone array, which comprises 32 capsules. However, arrays with far fewer capsules, such as tetrahedral arrays with four capsules, are much more widely available and employed in practice. Additionally, for SDM, the 3D intensity probe comprising six omnidirectional microphones has been more commonly used. Therefore, for the third and final test, the proposed reconstruction-based approach was compared against first-order SIRR and SDM when using these more accessible and affordable microphone arrays as input. The first-order SIRR and SDM test cases were rendered using the default configurations of the HO-SIRR [57] and SDM [20] MATLAB toolboxes, respectively. As an additional control, a first-order SH receiver was included, representing a low channel count receiver, but one that is free from microphone array design limitations. Note that the tetrahedral array was encoded into the Ambisonics format using the encoder described by Eq. (7), prior to applying the SIRR method. The test cases for the third part of the experiment are summarized in Table 3.

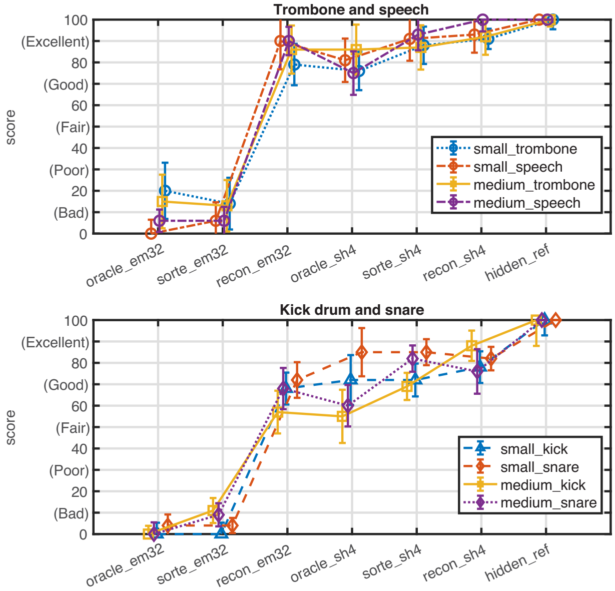

Fig. 3. Listening experiment part 1 results displaying medians and 95% confidence intervals.

6 RESULTS AND DISCUSSION

The results for the first part of the perceptual study are presented in Fig. 3. Note that the results for the more stationary source stimuli (trombone and speech) are separated from the more transient source stimuli (kick drum and snare) for improved visual clarity. For the cases involv-

Table 2. Test cases for listening experiment part 2.

Name	Receiver	Algorithm
hidden_ref	Quantized to t-design	Direct binauralization of quantized image sources
repair_sh4	Fourth-order SH receiver	Presented framework using the spatial reconstruction-based optimization
repair_em32	32-sensor rigid SMA	Presented framework using the spatial reconstruction-based optimization
hosirr_sh4	Fourth-order SH receiver	Rendered by the HO-SIRR toolbox
hosirr_em32_sh4	32-sensor rigid SMA	Encoded into fourth-order SH and rendered by the HO-SIRR toolbox

Table 3. Test cases for listening experiment part 3.

Name	Receiver	Algorithm
hidden_ref	Quantized to t-design	Direct binauralization of quantized image sources
repair_sh1	First-order SH receiver	Presented framework using the spatial reconstruction-based optimization
repair_tetra	Tetrahedral SMA	Presented framework using the spatial reconstruction-based optimization
repair_ip	Six-sensor open SMA	Presented framework using the spatial reconstruction-based optimization
sirr_sh1	First-order SH receiver	Rendered by the HO-SIRR toolbox
sirr_tetra_sh1	Tetrahedral SMA	Encoded into first-order SH and rendered by the HO-SIRR toolbox
sdm_ip	Six-sensor open SMA	Rendered by the SDM toolbox

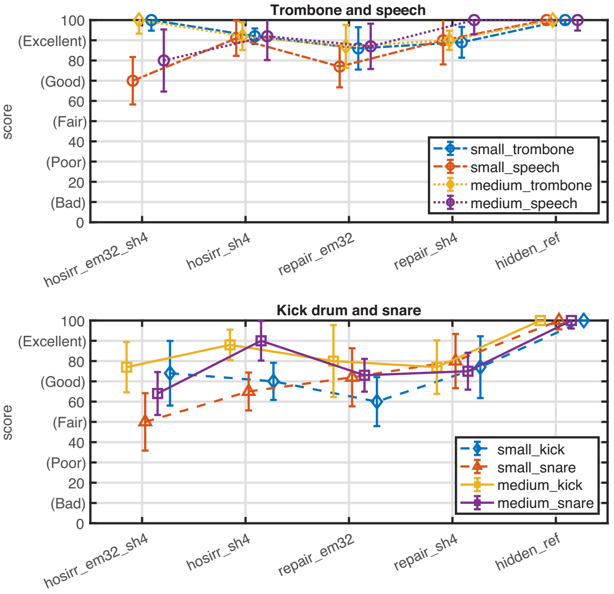

Fig. 4. Part 2 results displaying medians and 95% confidence intervals.

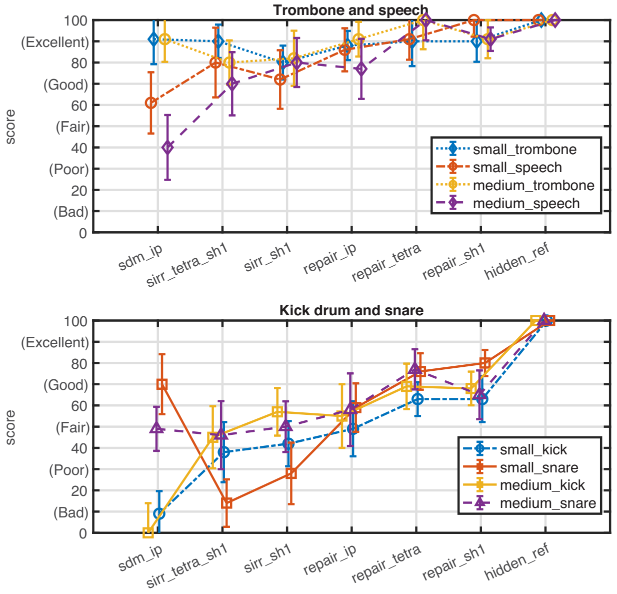

Fig. 5. Part 3 results displaying medians and 95% confidence intervals.

ing the fourth-order ideal SH receiver (*sh4*), the proposed reconstruction-based reflection number estimation scored similarly or higher than the SORTE and Oracle cases. However, when employing the Eigenmike32 configuration, the proposed approach was rated significantly higher than when SORTE or Oracle data were used. This demonstrates the effect of limitations in the employed rendering techniques when using imperfect microphone arrays. It suggests that even if the spatial analysis perfectly parameterizes the response, other aspects of the rendering (such as the use of beamformers with limited spatial selectivity) may result in an inferior perceptual outcome. However, by taking into consideration more aspects of the rendering architecture, the proposed optimization may find an alternative parameterization, which leads to an improved perceptual result.

The second experiment compared HO-SIRR against the presented parametric rendering framework, with the latter configured to use the proposed spatial reconstruction-based approach to estimate the number of reflections. The results are provided in Fig. 4. In this test, both methods obtained median ratings within the range denoted with the “Good” and “Excellent” verbal anchors in almost all cases. The one exception was the HO-SIRR method using the em32 receiver in conjunction with the snare stimulus and the smaller room. For both HO-SIRR and the proposed approach, the use of an ideal fourth-order SH receiver resulted in slightly higher ratings compared with using the em32 receiver. The ideal SH and em32 input were then rated similarly and toward the upper end of the scale when the responses were convolved with the more stationary (less transient) source material. This also largely extended to the cases in which the responses were convolved with transient source material, but with the proposed approach being rated slightly lower or higher than the HO-SIRR method depending on the source stimuli and room acoustics. However, because all of the scores are generally within the upper range of the evaluation scale, it may be concluded that both HO-SIRR and the proposed approach perform well when using receivers comprising a high number of sensors/channels.

The third experiment, which employed microphone arrays comprising far fewer sensors, demonstrated more variability in the results. These results are presented in Fig. 5. For this experiment, there was a stronger dependence on the source stimulus used. For example, when using the trombone stimulus, the proposed approach, SIRR, and SDM renders were all rated similarly and largely in the region labeled with the “Excellent” verbal anchor. However, convolving the rendered responses with the more transient stimuli resulted in lower scores across all of the tested methods. This suggests that impulsive material may be more revealing of artefacts incurred during the rendering, especially when using these more limited microphone array configurations. The transient kick drum then represented a particularly problematic signal for SDM. These issues likely arise because of the sparsity of the reproduced response, which has been studied as a general phenomenon in [58], and has prompted recent proposals for improvements in [59] and [60]. All in all, the test cases involving the use of the proposed reconstruction-based optimization tend to be among the best performing. The worst performance for the proposed approach was found when using the simulated intensity probe. However, when using this open array of omnidirectional sensors, the authors postulate that the limited performance of the beamformers used to extract the direct and residual components may have contributed to this apparent reduction in perceived accuracy. Despite this, however, the results were generally still higher than SDM, which is a method specifically intended for such arrays.

7 CONCLUSIONS AND FUTURE WORK

This article proposes a brute-force approach for the task of optimally rendering microphone array RIRs for auralization purposes. The employed rendering framework involves decomposing the input microphone array RIR (or Ambisonic RIR) into its individual reflections and anisotropic diffuse reverberation. The proposed approach then operates by first applying several different spatial analysis and/or reproduction strategies independently over time and frequency. This is then followed by determining which combination of approaches results in the most optimal reconstruction of the original RIR. Once the optimal rendering configuration has been established independently across time and frequency for the entire response, the microphone array transfer functions (used for the reconstruction) are substituted with spatialization gains corresponding to the target playback setup.

In this study, the proposed reconstruction-based approach was configured to determine the optimal number of reflections to render, since this was deemed by the authors as being the most significant aspect of the rendering framework that would affect its overall performance. A formal listening experiment was then conducted based on simulated RIRs. All test cases for the first part of the experiment employed the presented rendering framework while using either an ideal fourth-order Ambisonic receiver or a rigid SMA comprising 32 capsules. The results demonstrated that the proposed spatial reconstruction-based optimizations led to similar or better results than the other test cases. However, the main finding was that, when using the 32-capsule microphone array, renders using the proposed optimization were rated higher than when using ground-truth reflection numbers. This suggests that having such information does not necessarily mean that the employed microphone array and rendering techniques are able to make effective use of it. The demonstrated low perceptual performance, obtained when using the ground-truth reflection numbers, is thought to arise from the limited spatial selectivity of the employed beamformers. Therefore, applying the rendering using a subset of the captured reflections, or relying primarily on diffuse rendering strategies in the later part of the response, may be more perceptually favorable in practice.

The second part of the perceptual experiment compared the proposed rendering approach against an existing multidirectional rendering method. Here, the same ideal fourth-order Ambisonic receiver and 32-capsule spherical array were used. It was demonstrated that both methods, when using both microphone arrays, were rated similarly and toward the upper end of the evaluation scale when the responses were convolved with the more stationary (less transient) stimuli. This high level of perceived accuracy also largely extended to when the responses were convolved with more transient stimuli, with the proposed approach being rated slightly lower or higher than this existing method, depending on the room acoustics. However, it is noted that both methods were largely assigned scores within a range denoted with “Good” and “Excellent” verbal anchors on the test interface. This suggests that either method, when coupled with a high spatial resolution receiver, may lead to a high degree of perceived accuracy.

The final listening experiment investigated the perceived performance when using more limited microphone arrays, since such arrays are more widely available in practice. These arrays were as follows: an open spherical arrangement of six omnidirectional microphones, an open tetrahedral array of four cardioid microphones, and an ideal first-order Ambisonics receiver. Again, for the more stationary source material, the proposed approach was largely rated similarly to renders obtained through existing methods. However, for the more transient source material, the proposed approach was rated higher than these existing methods in the majority of cases.

The results of the listening experiments suggest that the proposed spatial reconstruction-based optimization has merit and may therefore warrant further investigations. Future work could involve characterizing its performance when using measured microphone array RIRs. One could also identify other aspects of the rendering framework for optimization, such as using different DoA estimators or beamformer designs. A key aspect of the proposed approach also lies in the choice of an appropriate cost-function, which is used for assessing the spatial reconstruction performance. Therefore, future work could involve investigating alternative cost-functions. Furthermore, one main downside of the proposed optimization is the inherent high level of computational complexity, since the rendering framework must be applied once per configuration considered. Although microphone array RIR rendering is typically conducted as an offline process, some applications may benefit from reduced computational complexity. This would naturally be achieved if fewer rendering configurations were to be considered for each time-frequency tile, for example, through the use of a more informed iterative scheme or by terminating the search when the cost-function error falls below a certain threshold or finds a local minimum. Finally, due to the general nature of the presented framework, it is noted that future work could also involve investigating its performance when using RIRs captured using non-SMAs, such as those integrated into mobile phones, 360-degree cameras, or head-worn devices, which could find application within future augmented reality contexts.

8 REFERENCES

[1] J. Pätynen, S. Tervo, and T. Lokki, "Analysis of Concert Hall Acoustics via Visualizations of Time-Frequency and Spatiotemporal Responses," *J. Acoust. Soc. Am.*, vol. 133, no. 2, pp. 842–857 (2013 Feb.). https://doi.org/10.1121/1.4770260.

[2] T. Lokki, J. Pätynen, A. Kuusinen, and S. Tervo, "Concert Hall Acoustics: Repertoire, Listening Position, and Individual Taste of the Listeners Influence the Qualitative Attributes and Preferences," *J. Acoust. Soc. Am.*, vol. 140, no. 1, pp. 551–562 (2016 Jul.). https://doi.org/10.1121/1.4958686.

[3] B. F. G. Katz and A. Weber, "An Acoustic Survey of the Cathédrale Notre-Dame de Paris Before and After the Fire of 2019," *Acoustics*, vol. 2, no. 4, pp. 791–802 (2020 Dec.). https://doi.org/10.3390/acoustics2040044.

[4] F. Melchior, C. Sladeczek, A. Partzsch, and S. Brix, "Design and Implementation of an Interactive Room Simulation for Wave Field Synthesis," in *Proceedings of the AES 40th International Conference: Spatial Audio: Sense the Sound of Space* (2010 Oct.), paper 7-5.

[5] A. Neidhardt and B. Reif, "Minimum BRIR Grid Resolution for Interactive Position Changes in Dynamic Binaural Synthesis," presented at the *148th Convention of the Audio Engineering Society* (2020 May), paper 10371.

[6] K. Müller and F. Zotter, "The PerspectiveLiberator–An Ipmixing 6DoF Rendering Plugin for Single-Perspective Ambisonic Room Impulse Responses," *arXiv preprint arXiv:2210.03360* (2022).

[7] T. Deppisch, S. Amengual Garí, P. Calamia, and J. Ahrens, "Perceptual Evaluation of Spatial Room Impulse Response Extrapolation by Direct and Residual Subspace Decomposition," in *Proceedings of the AES International Audio for Virtual and Augmented Reality Conference* (2022 Aug.), paper 14.

[8] J. Merimaa, "Spatial Impulse Response Rendering I: Analysis and Synthesis," *J. Audio Eng. Soc.*, vol. 53, nos. 12, pp. 1115–1127 (2005 Dec.).

[9] S. Tervo, J. P. Tynen, A. Kuusinen, and T. Lokki, "Spatial Decomposition Method for Room Impulse Responses," *J. Audio Eng. Soc.*, vol. 61, nos. 1–2, pp. 17–28 (2013 Jan).

[10] Y. Yamasaki and T. Itow, "Measurement of Spatial Information in Sound Fields by Closely Located Four Point Microphone Method," *J. Acoust. Soc. Jpn. (E)*, vol. 10, no. 2, pp. 101–110 (1989). https://doi.org/10.1250/ast.10.101.

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

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

[13] 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. 368–354 (2020 May). https://doi.org/10.17743/jaes.2020.0026.

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

[15] L. McCormack, S. Delikaris-Manias, A. Politis,Â et al., â€œApplications of Spatially Localized Active-Intensity Vectors for Sound-Field Visualization," *J. Audio Eng. Soc.*, vol. 67, no. 11, pp. 840–854 (2019 Nov.). https://doi.org/10.17743/jaes.2019.0041.

[16] A. Politis, J. Vilkamo, and V. Pulkki, "Sector-Based Parametric Sound Field Reproduction in the Spherical Harmonic Domain," *IEEE J. Sel. Top. Signal Process.*, vol. 9, no. 5, pp. 852–866 (2015 Aug.). https://doi.org/10.1109/JSTSP.2015.2415762.

[17] A. Politis, L. McCormack, and V. Pulkki, "Enhancement of Ambisonic Binaural Reproduction Using Directional Audio Coding With Optimal Adaptive Mixing," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 379–383 (New Paltz, NY) (2017 Oct.).

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

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

[20] S. Tervo, "SDM Toolbox (Version 1.3001)," https://se.mathworks.com/matlabcentral/fileexchange/56663-sdm-toolbox (accessed March 25, 2023).

[21] N. Meyer-Kahlen, S. J. Schlecht, and T. Lokki, "Parametric Late Reverberation From Broadband Directional Estimates," in *Proceedings of International Conference on Immersive and 3D Audio: From Architecture to Automotive (I3DA)* (Bologna, Italy) (2021 Sep.).

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

[23] P. Coleman, A. Franck, P. J. B. Jackson, R. J. Hughes, L. Remaggi, and F. Melchior, "Object-Based Reverberation for Spatial Audio," *J. Audio Eng. Soc.*, vol. 65, nos. 1–2, pp. 66–77 (2017 Jan.). https://doi.org/10.17743/jaes.2016.0059.

[24] P. Stade, J. Arend, and C. Pörschmann, "A Parametric Model for the Synthesis of Binaural Room Impulse Responses," *Proc. Mtgs. Acoust.* vol. 30 , no. 1 , paper 015006 (2017 Jun.). https://doi.org/10.1121/2.0000573.

[25] L. Gölles and F. Zotter, "Directional Enhancement of First-Order Ambisonic Room Impulse Responses by the 2+2 Directional Signal Estimator," in *Proceedings of the*

15th International Audio Mostly Conference, pp. 38–45 (Graz. Austria) (2020 Sep.).

[26] V. Gunnarsson and M. Sternad, "Binaural Auralization of Microphone Array Room Impulse Responses Using Causal Wiener Filtering," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 2899–2914 (2021 Sep.). https://doi.org/10.1109/TASLP.2021.3110340.

[27] S. Tervo and A. Politis, "Direction of Arrival Estimation of Reflections from Room Impulse Responses Using a Spherical Microphone Array," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 10, pp. 1539–1551 (2015 Oct.). https://doi.org/10.1109/TASLP.2015.2439573.

[28] R. Schmidt, "Multiple Emitter Location and Signal Parameter Estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280 (1986 Mar.).

[29] T. Deppisch, S. V. A. Garí, P. Calamia, and J. Ahrens, "Direct and Residual Subspace Decomposition of Spatial Room Impulse Responses," *arXiv preprint arXiv:2207.09733* (2022).

[30] A. Politis, S. Tervo, and V. Pulkki, "COMPASS: Coding and Multidirectional Parameterization of Ambisonic Sound Scenes," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6802–6806 (Calgary, Canada) (2018 Apr.).

[31] L. McCormack, A. Politis, R. Gonzalez, T. Lokki, and V. Pulkki, "Parametric Ambisonic Encoding of Arbitrary Microphone Arrays," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 30 , 2062–2075 (2022 Jun.). https://doi.org/10.1109/TASLP.2022.3182857.

[32] L. McCormack, N. Meyer-Kahlen, and A. Politis, "Multi-directional Parameterisation and Rendering of Spatial Room Impulse Responses," in *Proceedings of the 24th International Congress on Acoustics (ICA)* (Gyeongju, Korea) (2022 Oct.).

[33] B. Rafaely, *Fundamentals of Spherical Array Processing*, vol. 8 (Springer Cham, Switzerland, 2015), 2nd ed. https://doi.org/10.1007/978-3-319-99561-8.

[34] E. G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography* (Academic Press, Cambridge, MA, 1999).

[35] H. Teutsch, *Modal Array Signal Processing: Principles and Applications of Acoustic Wavefield Decomposition* (Springer Berlin, Heidelberg, Germany, 2007).

[36] S. Moreau, J. Daniel, and S. Bertet, "3D Sound Field Recording With Higher Order Ambisonics–Objective Measurements and Validation of a 4th Order Spherical Microphone," presented at the *120th Convention of the Audio Engineering Society* (2006 May), paper 6857.

[37] T. Deppisch, J. Ahrens, S. V. A. Garí, and P. Calamia, "Spatial Subtraction of Reflections from Room Impulse Responses Measured with a Spherical Microphone Array," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 346–350 (New Paltz, NY) (2021 Oct.).

[38] A. Politis and H. Gamper, "Comparing Modeled and Measurement-Based Spherical Harmonic Encoding filters for Spherical Microphone Arrays," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 224–228 (New Paltz, NY) (2017 Oct.).

[39] C. T. Jin, N. Epain, and A. Parthy, "Design, Optimization and Evaluation of a Dual-Radius Spherical Microphone Array," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 22, no. 1, pp. 193–204 (2014 Jan.). https://doi.org/10.1109/TASLP.2013.2286920.

[40] M. A. Doron and A. Nevet, "Robust Wavefield Interpolation for Adaptive Wideband Beamforming," *Signal Process.*, vol. 88, no. 6, pp. 1579–1594 (2008 Jun.).

[41] N. Huleihel and B. Rafaely, "Spherical Array Processing for Acoustic Analysis Using Room Impulse Responses and Time-Domain Smoothing," *J. Acoust. Soc. Am.*, vol. 133, no. 6, pp. 3995–4007 (2013 Jun.). https://doi.org/10.1121/1.4804314.

[42] T.-J. Shan, M. Wax, and T. Kailath, "On Spatial Smoothing for Direction-of-Arrival Estimation of Coherent Signals," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 33, no. 4, pp. 806–811 (1985 Aug.).

[43] H. Beit-On and B. Rafaely, "Focusing and Frequency Smoothing for Arbitrary Arrays with Application to Speaker Localization," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 2184–2193 (2020 Jul.). https://doi.org/10.1109/TASLP.2020.3010098.

[44] L. McCormack, R. Gonzalez, J. Fernandez, C. Hold, and A. Politis, "Parametric Ambisonic Encoding using a Microphone Array With a One-Plus-Three Configuration," in *Proceedings of the AES International Audio Conference for Virtual and Augmented Reality* (2022 Aug.), paper 16.

[45] K. Han and A. Nehorai, "Improved Source Number Detection and Direction Estimation With Nested Arrays and ULAs Using Jackknifing," *IEEE Trans. Signal Process.*, vol. 61, no. 23, pp. 6118–6128 (2013 Dec.). https://doi.org/10.1109/TSP.2013.2283462.

[46] W. Chen, K. M. Wong, and J. P. Reilly, "Detection of the Number of Signals: A Predicted Eigen-Threshold Approach," *IEEE Trans. Signal Process.*, vol. 39, no. 5, pp. 1088–1098 (1991 May).

[47] J.-S. Jiang and M. A. Ingram, "Robust Detection of Number of Sources Using the Transformed Rotational Matrix," in *Proceedings of the IEEE Wireless Communications and Networking Conference*, vol. 1, pp. 501–506 (Atlanta, GA) (2004).

[48] H. Cox, R. Zeskind, and T. Kooij, "Practical Supergain," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 34, no. 3, pp. 393–398 (1986 Jun.).

[49] M.-V. Laitinen, J. Vilkamo, K. Jussila, A. Politis, and V. Pulkki, "Gain Normalization in Amplitude Panning as a Function of Frequency and Room Reverberance," in *Proceedings of the AES 55th International Conference: Spatial Audio* (2014 Aug.), paper 3-5.

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

[51] J. Vilkamo, T. Bäckström, and A. Kuntz, "Optimized Covariance Domain Framework for Time–Frequency Processing of Spatial Audio," *J. Audio Eng. Soc.*, vol. 61, no. 6, pp. 403–411 (2013 Jun.).

[52] L. McCormack and A. Politis, "Estimating and Reproducing Ambience in Ambisonic Recordings," in *Proceedings of the 30th European Signal Processing Conference (EUSIPCO)*, pp. 314–318 (Belgrade, Serbia) (2022 Aug./Sep.).

[53] J. B. Allen and D. A. Berkley, "Image Method for Efficiently Simulating Small-Room Acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950 (1979 Apr.).

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

[55] C. Armstrong, L. Thresh, D. Murphy, and G. Kearney, "A Perceptual Evaluation of Individual and Non-individual HRTFs: A Case Study of the SADIE II Database," *Appl. Sci.*, vol. 8, no. 11, paper 2029 (2018 Nov.).

[56] J. Meyer and G. Elko, "A Highly Scalable Spherical Microphone Array Based on an Orthonormal Decomposition of the Soundfield," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, pp. 1781–1784 (Orlando, FL) (2002 May).

[57] L. McCormack, A. Politis, O. Scheuregger, and V. Pulkki, "Higher-Order Processing of Spatial Impulse Responses," in *Proceedings of the 23rd International Congress on Acoustics (ICA),* pp. 4909–4916 (Aachen, Germany) (2019 Sep.).

[58] N. Meyer-Kahlen, S. J. Schlecht, and T. Lokki, "Perceptual Roughness of Spatially Assigned Sparse Noise for Rendering Reverberation," *J. Acoust. Soc. Am.*, vol. 150, no. 5, pp. 3521–3531 (2021 Nov.). https://doi.org/10.1121/10.0007048.

[59] 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 Dec.). https://doi.org/10.17743/jaes.2020.0063.

[60] 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 (2022 Dec.). https://doi.org/10.17743/jaes.2022.0039.

THE AUTHORS

Leo McCormack

Nils Meyer-Kahlen

Archontis Politis

Leo McCormack is a doctoral candidate for the Department of Signal Processing and Acoustics at Aalto University, Finland, researching parametric spatial audio technologies. He received his M.Sc. degree in Computer Communications and Information Sciences, majoring in Acoustics and Audio Technology, at Aalto University, Finland, and his B.Sc. in Music Technology and Audio Systems at the University of Huddersfield, UK. His research interests include multichannel and microphone array signal processing for sound field reproduction.

•

Nils Meyer-Kahlen is a doctoral candidate for the Department of Signal Processing and Acoustics at Aalto University in Finland. Before joining the lab in 2019, he completed his B.Sc. and M.Sc. in Electrical Engineering and Audio Engineering at the Technical University and the University of Music and Performing Arts in Graz, Austria. His main research interest is virtual acoustics for augmented reality, from both a technological and a perceptual point of view.

•

Archontis Politis is a post-doctoral researcher at Tampere University, Finland. He obtained his M.Sc. degree in Sound and Vibration studies from the Institute of Sound and Vibration Research (ISVR), University of Southampton, UK, in 2008. In 2015, he was a visiting researcher at the University of Maryland Institute for Advanced Computer Studies, Maryland, USA, and in the same year, he completed a research internship at Microsoft Research, Redmond, Washington, USA. In 2016, he obtained a Doctor of Science degree on spatial audio processing from Aalto University, Finland. He has served as editor of a book on Parametric Spatial Audio Processing and an organizer in the DCASE scientific challenge and has chaired various special sessions in international conferences. His research interests include spatial audio technologies, virtual acoustics, array signal processing, and acoustic scene analysis.