Greco *et al.*
EURASIP Journal on Audio, Speech, and Music Processing (2026) 2026:4
https://doi.org/10.1186/s13636-025-00437-y

EURASIP Journal on Audio, Speech, and Music Processing

METHODOLOGY

Open Access

Parametric virtual microphone techniques for sound field reconstruction with early reflection modeling

Gioele Greco[1*], Mirco Pezzoli[1], Fabio Antonacci[1] and Augusto Sarti[1]

Abstract

Recent advances in immersive media and virtual environments have highlighted the crucial role of spatial audio in enhancing perceptual realism, enabling true six-degrees-of-freedom experiences in applications such as virtual reality, augmented reality, and advanced teleconferencing. Nonetheless, accurately reconstructing the direct sound, diffuse field, and early reflections, crucial for spatial depth and realism, remains challenging, especially with a limited number of measurements. To address this challenge, we introduce Parametric modeling of Direct sound, Early reflections, and Reverberation (ParaDER), a unified parametric framework that explicitly separates and reconstructs direct sound, early reflections, and diffuse reverberation from a minimal set of measurements. In the first stage, sound sources are localized by solving a sparse, regularized optimization problem that yields low-order spherical-harmonic coefficients and thus captures the direct component of the field. The second stage follows image-source theory: the estimated room impulse response is segmented into a small number of early reflections, each of which is modeled as an image source whose position and amplitude are fitted to the segmented data. This explicit treatment preserves the temporal and spatial characteristics of early reflections, which are critical for accurate depth perception and localization cues. In the final stage, the estimated direct and early reflection components are analytically propagated to virtual microphone positions, and the remaining energy is synthesized as diffuse reverberation under an isotropic assumption. Because the entire pipeline is low-order, comprising only a few source and image-source parameters, ParaDER can reconstruct a spatial sound field with few physical microphones, reducing memory and computation compared to both other parametric methods and non-parametric approaches. Extensive evaluations in 100 simulated shoebox rooms confirm that ParaDER markedly improves reconstruction accuracy. When compared to a state of the art parametric model, the normalized mean-squared error of the acoustic metrics shows that the estimate of the early reflections improves the accuracy. Also, subjective listening tests on a real conference room dataset, show that our method yields higher mean MUSHRA scores for both speech and music. Listeners consistently report clearer spatial cues, more precise localization, and a more natural timbre, demonstrating that explicit modeling of early reflections confers perceptually significant benefits.

Keywords Spatial audio, Immersive audio, Virtual microphone, Early reflections, Acoustic parametric modeling

1 Introduction

Recent advances in immersive media and virtual environments have underscored the main role of spatial audio in enhancing perceptual realism. Immersive audio systems enable interactive sound experiences through free head and body movements, known as six-degrees of freedom (6DoF), offering listeners the unique capability

*Correspondence:
Gioele Greco
gioele.greco@polimi.it
[1] Dipartimento Elettronica Informatica Bioingegneria, Politecnico di Milano, Via Ponzio 34/5, 20133 Milano, Italy

to experience sound fields that adapt naturally to their position and orientation. Such features are increasingly required in applications such as virtual reality (VR), augmented reality (AR), and advanced teleconferencing [1–3].

With appropriate analysis and synthesis techniques, this enables a listener to navigate a recorded acoustic scene while perceiving spatial sound characteristics as if he/she were in the virtual recording location [4]. The reconstruction of a sound field is a well-known and challenging problem in the acoustic signal processing community, often referred to as the virtual miking or sound field estimation problem. This process aims to reconstruct the signal that would be captured by a virtual microphone (VM) arbitrarily placed in space, based on signals acquired from an array of physical microphones.

Despite significant advancements in 6DoF sound field estimation, accurately reconstructing early reflections (ER), those that shape spatial perception and depth, remains a challenging task [5, 6]. While some recent methods [7–9] have begun to incorporate ER estimation using simplified geometric or parametric models, they often rely on strong idealized assumptions. As a result, these approaches struggle to preserve the full range of perceptual cues conveyed by ERs when applied to more complex or arbitrary acoustic environments. ERs play a critical role in conveying information about the acoustic space, significantly affecting both spatial localization and the perceived realism of a sound scene. However, as it will be discussed, existing methods either ignore these reflections altogether or approximate them so coarsely that many of their perceptual advantages are lost.

This work aims to address that gap by developing a unified framework that not only reconstructs the 6DoF sound field but also delivers a high-fidelity estimation of ER.

Existing sound field reconstruction approaches can be broadly classified into non-parametric and parametric methods. Parametric methods [10–17, 19] model the sound field in a compact manner by estimating spatial and physically meaningful parameters, which can then be used to synthesize the desired sound signals. In contrast, non-parametric methods [20–28] aim to reconstruct the sound field directly from measurements without explicitly estimating scene parameters such as source positions or room geometry.

Early methods [13, 29, 30] relied on Plane Wave Decomposition and spatial Fourier transforms, using basis functions derived from the wave equation. In [30], a compact parameterization of the acoustic transfer function in reverberant environments was introduced, while [29] extended this concept to spatial harmonic coefficient translation for multi-zone sound field reproduction.

Higher-order representations, such as Spherical Harmonic Decomposition (SHD), have been widely used in Higher-Order Ambisonics (HOA) for sound field reconstruction [31–33]. These techniques allow for compact encoding of spatial information.

However, despite offering solid theoretical grounding, both plane wave and spherical harmonic-based approaches require dense microphone arrays to avoid spatial aliasing. Their accuracy depends heavily on the number and distribution of available microphones. In cases where the sampling density is low or the environment is unknown, reconstruction errors become significant due to missing spatial information, limiting the practicality of these methods in sparse measurement scenarios.

To mitigate the need for dense microphone distributions, researchers have explored compressive sensing (CS) [8, 34] and sparse representations [25, 36, 37]. In [34] authors demonstrated that CS combined with plane wave decomposition could reconstruct sound fields using a reduced number of measurements. Similarly, [25] proposed a sparse representation framework to approximate the acoustic transfer function with minimal data. Moreover, in [8], they tried to improve the model proposed in [25] to estimate ER based on the knowledge of the room geometry. However, these approaches face high-frequency reconstruction limitations due to the fact that compressive sensing assumes sparsity, which does not always hold in complex, reverberant environments.

In [25, 38], hybrid models are proposed that explicitly account for reverberation effects, leading to improved reconstruction of ER. These models enhance the accuracy of direct sound field estimation; however, they also come with increased computational demands. As noted in [25], such demands are related to the inclusion of additional hyperparameters and the need to represent a more complex set of variables.

Furthermore, the method struggles in environments with strong anisotropic reflections, where traditional assumptions of diffuse reverberation fail. Such directional reflections degrade the separation performance, requiring more intricate algorithms and higher processing power to achieve robust sound field reconstruction.

Recent advancements in machine learning have opened new possibilities for sound field reconstruction in unknown spaces. The work in [39] employed Gaussian Process Regression (GPR) with anisotropic kernels to model directional sound propagation, achieving accurate reconstructions over large spatial domains. Moreover, [40, 41] proposed the use of complex-valued neural networks and diffusion models to estimate room transfer functions from a limited number of microphones in

specific environments. While effective under controlled conditions, these methods demand significant computational resources and are prone to overfitting when training data is scarce. Furthermore, they rely on accurate knowledge of the room transfer functions at the microphone positions, which is rarely available in real-world scenarios.

A promising direction to address some of these limitations is the use of Physics-Informed Neural Networks (PINNs), which embed wave propagation models directly into the training process [42, 43]. PINNs can reduce the dependence on large datasets and have demonstrated the ability to generalize to unseen environments by incorporating physical constraints. Nevertheless, their generalization capabilities beyond the training domain remain uncertain. Additionally, PINNs are computationally intensive and require carefully designed loss functions to enforce physically consistent behavior.

Due to these limitations, particularly the challenges in generalization and data availability, we ultimately favor parametric models, which offer more interpretable and robust performance across varying acoustic conditions.

In contrast to non-parametric approaches, parametric methods infer underlying scene properties, such as source positions, signals, and room responses, allowing for targeted reconstruction based on these parameters. Previous works [31–33] have demonstrated that a Higher Order Microphone (HOM) can represent the sound field in the SHD. This transformation provides an accurate depiction of the local sound field, which is easily converted into a 3DoF Head-Related Transfer Function. Our approach focuses on estimating the n-th order expansion at a virtual HOM center using signals from spatially distributed HOMs. As previously discussed, some methods, such as those presented in [8, 35], address the problem by also considering the ER. However, these approaches require detailed information about the room geometry, which is often unknown or difficult to obtain. In contrast, our method aims to reconstruct the sound field without the need for explicit knowledge of the room geometry. This is crucial for spaces where geometries are complex or inaccessible.

A fast spatial interpolation method was proposed in [9], relying solely on microphone positions and assuming a spatially isotropic sound field. The approach estimates simple spatial parameters in the time-frequency domain that summarize the acoustic scene at the microphone locations, enabling audio reconstruction up to first-order Ambisonics. While versatile, this assumption fails in scenarios with directional sources or ER, both critical for spatial perception fidelity.

Other models [44] decompose the sound field into its direct and diffuse components. The direct component, originating from the sound sources, is represented using SHD. In the analysis phase, the SHD coefficients are obtained by solving a sparse, regularized optimization problem.

This decomposition necessitates estimating source parameters such as locations, signals, and directivity patterns. These estimations can be achieved through techniques like source localization using distributed microphone arrays [45–47], source separation methods [48], and directivity estimation strategies.

The diffuse component, conversely, is assumed to be isotropic and homogeneous, and it is estimated from the analysis of signals at at least one microphone pair within the HOM. However, estimations from different microphone pairs could be different, due to the residual direct-component contamination. To mitigate this, and similarly to [15], the diffuse component at the VM is calculated as a weighted sum of estimations from all the microphone pairs.

Despite the robustness of this direct–diffuse split, these approaches typically ignore ER, which are critical for conveying spatial depth and realism in a reconstructed sound field. This omission constitutes a fundamental limitation when aiming at a fully immersive audio reproduction.

Motivated by these prior approaches, our objective is to bridge the existing performance gap and enable highly accurate sound field reconstruction using only a limited number of HOM measurements. To this end, we explicitly integrate early-reflection estimation into our framework. Specifically, we developed Parametric modeling of Direct sound, Early reflections and Reverberation (ParaDER) to incorporate the estimation of virtual source parameters by leveraging the image source model, and we account for non-idealities by estimating a filter function that characterizes the acoustic transformation associated with each virtual source.

The paper is structured as follows: in Sect. 2, we introduce the sound field data model and present the problem formulation, establishing the theoretical framework and key assumptions. Section 3 describes the strategies for retrieving the model parameters. In Sect. 4, we focus on the synthesis of the sound field at the VM location. Section 5 reviews our experimental outcomes, comparing simulated results with those from perceptual tests in a virtual 6DoF environment using real measurements from [49]. Finally, Sect. 6 draws the conclusions.

1.1 Notation

To aid the reader with the notation used in this document, we provide here a table collecting all the useful symbols meaning (Table 1).

Table 1 Notation conventions used throughout the document

Symbol	Application	Meaning
$'$	Superscript	Source-related variable
$''$	Superscript	Virtual source-related variable
$\hat{}$	Superscript	Estimated parameter
$\check{}$	Superscript	Virtual microphone-related variable
n	Subscript	Source index
i	Subscript	Virtual source index
a	Subscript	Microphone array index
q	Subscript	Microphone capsule index
v	Subscript	Virtual microphone index

Combination examples:

- $\boldsymbol{r}''_{i,n}$: Position of the i-th virtual source associated with the n-th physical source
- $\boldsymbol{r}''_{i,a,n}(t)$: Position of the i-th virtual source (associated with the n-th physical source) captured by the a-th array
- $\hat{\boldsymbol{r}}'_{v}$: Estimated position of the v-th virtual microphone relative to a physical source

2 Data model and problem formulation

This section presents an overview of the data model along with the virtual miking problem. We begin by introducing the data model in Sec. 2.1, where we describe how the microphones capture the acoustic scene. Additionally, we define the adopted VM framework and describe the sound field parameters that require estimation. Following this, Sec. 2.2 details the virtual miking problem. Using a block diagram, we illustrate the fundamental functional blocks of the proposed approach.

2.1 Data model

Consider a Cartesian coordinate system where N acoustic sources are positioned at arbitrary locations $\boldsymbol{r}'_n = [x'_n, y'_n, z'_n]^T$, for $n = 1, \ldots, N$; a network of $A \geq 2$ distributed compact microphone arrays, each with M microphones, located at $\boldsymbol{r}_q = [x_q, y_q, z_q]^T$, for $q = 1, \ldots, M \times A$; and a set of V VMs (VMs) placed at $\check{\boldsymbol{r}}_v = [\check{x}_v, \check{y}_v, \check{z}_v]^T$, for $v = 1, \ldots, V$, as illustrated in Fig. 1.

Let us consider the measured data on qth microphone placed in $\boldsymbol{r}_q$. Assuming the N source signals are sufficiently sparse in the time-frequency domain [11, 15], so that a single source dominates each time-frequency bin, and considering that the signal is affected by noise $\Theta(t, \omega, \boldsymbol{r}_q)$, we model the acquired signal as a linear combination of an early sound component and a diffuse sound component, generalizing [15] as

$$X(t, \omega, \boldsymbol{r}_q) = C_q(\omega) X_{n,\text{early}}(t, \omega, \boldsymbol{r}_q) + D_q(\omega) X_{\text{diff}}(t, \omega, \boldsymbol{r}_q) + \Theta(t, \omega, \boldsymbol{r}_q), \tag{1}$$

where t is the time-frame index, $\omega = 2\pi f$ is the radial frequency with $f > 0$ as the temporal frequency, $C_q(\omega) \in \mathbb{C}$ models the qth microphone pick-up pattern, and $D_q(\omega) \in \mathbb{C}$ represents its sensitivity to the diffuse field. The term $X_{n,\text{early}}(t, \omega, \boldsymbol{r}_q)$ represents the direct sound emitted by the nth source and received by the qth microphone along with the components strongly correlated with the source signal, i.e., the so-called ER. The term $X_{\text{diff}}(t, \omega, \boldsymbol{r}_q)$ represents the diffuse sound field component and it is assumed to be spatially isotropic

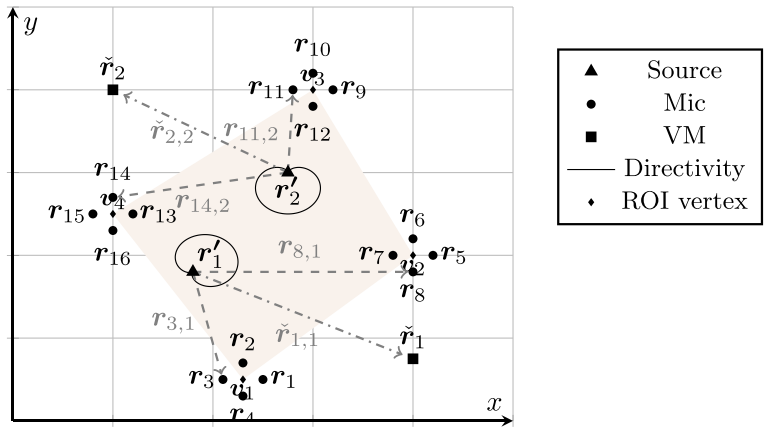

Fig. 1 Graphical representation of the 2D model. The setup consists of $A = 4$ circular microphone arrays, each containing $M = 4$ microphones. Two sources ($N = 2$) and two VMs ($V = 2$) are included in the scene. The source directivity function is overlaid on the scene diagram

and homogeneous, namely, it arrives with equal strength from all directions and its mean power does not vary with the position [15, 50].

The term $\Theta(t,\omega,\boldsymbol{r}_q)$ represents the additive noise of the qth sensor, respectively, modeled as uncorrelated, zero-mean complex Gaussian noise with mean power

$$\Phi_{N,qq}(t,\omega) = E\{\Theta((t,\omega,\boldsymbol{r}_q)\Theta(^{\star}(t,\omega,\boldsymbol{r}_q)\}, \tag{2}$$

where $E\{\cdot\}$ denotes the mathematical expectation and $(\cdot)^{\star}$ refers to the conjugate of a complex number.

A key novelty of our approach lies in the explicit inclusion of ER in the signal model. In particular, the term $X_{n,\text{early}}(t,\omega,\boldsymbol{r}_q)$ encompasses not only the direct sound emitted by the nth source and received by the qth microphone, but also the sound field component that is strongly coherent with the source signal and not spatially isotropic.

To provide a clear description of the term $X_{n,\text{early}}(t,\omega,\boldsymbol{r}_q)$ in (1) we outline the system geometry in spherical coordinates. Let us denote with $\boldsymbol{r}_{q,n} = \boldsymbol{r}_q - \boldsymbol{r}'_n = \left[x_{q,n}, y_{q,n}, z_{q,n}\right]^T$ the vector pointing from the source position to the microphone position (see Fig. 1) and with $\rho_{q,n}$, $\theta_{q,n}$ and $\phi_{q,n}$ the coordinates of $\boldsymbol{r}_{q,n}$ in a spherical coordinate system, i.e.,

$$\begin{aligned}
\rho_{q,n} &= \sqrt{x_{q,n}^2 + y_{q,n}^2 + z_{q,n}^2},\\
\theta_{q,n} &= \arccos\frac{\mathrm{z_{q,n}}}{\rho_{\mathrm{q,n}}},\\
\phi_{q,n} &= \arctan\frac{\mathrm{y_{q,n}}}{\mathrm{x_{q,n}}}.
\end{aligned} \tag{3}$$

The early part of the sound field is modeled as

$$X_{n,\text{early}}(t,\omega,\boldsymbol{r}_q) = X_{n,\text{direct}}(t,\omega,\boldsymbol{r}_q) + X_{n,\text{ER}}(t,\omega,\boldsymbol{r}_q), \tag{4}$$

where $X_{n,\text{direct}}$ represents the "direct" sound determined by the propagation through the direct path between the source and the qth receiver and $X_{n,\text{ER}}(t,\omega,\boldsymbol{r}_q)$ models the ER which are coherent with the the direct component. In order to take into account for possible directional properties of the sound source, we describe the direct sound though a spherical harmonics expansion as [44]

$$X_{n,\text{direct}}(t,\omega,\boldsymbol{r}_q) = \sum_{\ell=0}^{L}\sum_{\mu=-\ell}^{\ell}\beta_{\ell\mu}^{n}(t,\omega)h_\ell\left(k\rho_{q,n}\right)Y_{\ell\mu}\left(\theta_{q,n},\phi_{q,n}\right), \tag{5}$$

where $k=\omega/c$, with c the speed of sound, $\beta_{\ell\mu}^{n}(\omega)$ are the exterior sound field coefficients of the nth source, $h_\ell(\cdot)$ is the ℓth order spherical Hankel function and $Y_{\ell\mu}\left(\theta_{q,n},\phi_{q,n}\right)$ is the spherical harmonic of order ℓ and degree μ, defined as

$$Y_{\ell\mu}\left(\theta_{q,n},\phi_{q,n}\right) = K_{\ell\mu}P_{\ell\mu}\left(\cos(\theta_{q,n})\right)e^{j\mu\phi_{q,n}}, \tag{6}$$

with

$$K_{\ell\mu} = (-1)^{\mu}\sqrt{\frac{(2\ell+1)}{4\pi}\frac{(\ell-\mu)!}{(\ell+\mu)!}}, \tag{7}$$

and $P_{\ell\mu}(\cdot)$ the normalized associated Legendre polynomial.

We model the component $X_{n,\text{ER}}(t,\omega,\boldsymbol{r}_q)$ exploiting the well-known image-source method [51]. In fact, ERs preserve coherence with the direct path of the source signal, as they are considered reflections of the source sound pressure wave off surfaces (e.g., room walls); thus, they do not satisfy the diffuse conditions.

In practice, for each source n present in the space, we assume the existence of I_n Virtual Sources (VSs), located at positions $\boldsymbol{r}''_{i,n}$ far enough from the source location $\boldsymbol{r}'_n$. In the absence of information about the room geometry, we define $H_{n,i}(t,\omega,\boldsymbol{r}_q)$ as a transfer function modeling the relationship between the ith source signal and the nth related VS with respect to. the qth microphone

$$H_{n,i}(t,\omega,\boldsymbol{r}_q) = \frac{\sum_{\ell=0}^{L}\sum_{\mu=-\ell}^{\ell}\beta_{\ell\mu}^{n,i}(t,\omega)h_\ell\left(k\rho_{q,i,n}\right)Y_{\ell\mu}\left(\theta_{q,i,n},\phi_{q,i,n}\right)}{\sum_{\ell=0}^{L}\sum_{\mu=-\ell}^{\ell}\beta_{\ell\mu}^{n}(t,\omega)h_\ell\left(k\rho_{q,n}\right)Y_{\ell\mu}\left(\theta_{q,n},\phi_{q,n}\right)}, \tag{8}$$

where $\rho_{q,i,n},\theta_{q,i,n},\phi_{q,i,n}$ are the ith VS coordinates, defined as in (3).

The ER component is then given as the summation of the all the I_n VSs

$$X_{n,\text{ER}}(t,\omega,\boldsymbol{r}_v) = \sum_{i}^{I_n}H_{n,i}(t,\omega,\boldsymbol{r}_v)\sum_{\ell=0}^{L}\sum_{\mu=-\ell}^{\ell}\beta_{\ell\mu}^{n}(t,\omega)h_\ell\left(k\rho_{q,n}\right)Y_{\ell\mu}\left(\theta_{q,n},\phi_{q,n}\right), \tag{9}$$

and, therefore, the early component in (1) $X_{n,\text{early}}(t,\omega,\boldsymbol{r}_q)$ is finally defined as

$$\begin{aligned}
X_{n,\text{early}}(t,\omega,\boldsymbol{r}_q) = &\sum_{\ell=0}^{L}\sum_{\mu=-\ell}^{\ell}\beta_{\ell\mu}^{n}(t,\omega)h_\ell\left(k\rho_{q,n}\right)Y_{\ell\mu}\left(\theta_{q,n},\phi_{q,n}\right)+\\
&+\sum_{\hat{n}}^{I_n}H_{n,i}(t,\omega,\boldsymbol{r}_q)\sum_{\ell=0}^{L}\sum_{\mu=-\ell}^{\ell}\beta_{\ell\mu}^{n}(t,\omega)h_\ell\left(k\rho_{q,n}\right)Y_{\ell\mu}\left(\theta_{q,n},\phi_{q,n}\right),
\end{aligned} \tag{10}$$

Assuming the known microphone signal model described in (1), we similarly define the signal sensed at the VM location $\check{\boldsymbol{r}}_v$ as

$$\begin{aligned}
S(t,\omega,\check{\boldsymbol{r}}_v) = C_v(\omega)\Big(&S_{n,\text{direct}}(t,\omega,\check{\boldsymbol{r}}_v) + S_{n,\text{ER}}(t,\omega,\check{\boldsymbol{r}}_v)\Big)\\
&+ D_v(\omega)S_{\text{diff}}(t,\omega,\check{\boldsymbol{r}}_v),
\end{aligned} \tag{11}$$

It is worth noticing that in (11), we implicitly assume that the VM is noiseless. This signal model allows us to describe the sound field at the VM location using only the parameters related to the active sources. In the next section, we discuss how to estimate these parameters

based solely on the information provided by the Q microphones placed in the scene.

2.2 Problem formulation

The data available to solve the estimation problem are the signals $X(t,\omega,\boldsymbol{r}_q)$, $q=1,\dots,Q$ of the microphones, their positions $\boldsymbol{r}_q$, the characteristics of each VM, namely the position $\check{\boldsymbol{r}}_v$, the pick-up pattern $C_v(\omega)$ and the sensitivity to diffuse noise $D_v(\omega)$ and the number of sources N. In particular, as regards the latter parameter, it can be estimated using other sensors in the room (e.g., videocamera) or directly from the signals at the microphones as proposed, for example, in [52–59]. The output of the algorithm is an estimate $\hat{S}(\check{\boldsymbol{r}}_v)$ of the VM signal $S(\check{\boldsymbol{r}}_v)$.

In this paper, we propose a modular framework for sound field reconstruction, specifically designed to offer flexibility in scenarios where certain information, such as source signals or positions, is already available. This adaptability allows users to bypass specific processing steps and focus computational resources where they are most needed. A graphical overview of the proposed architecture is provided in Fig. 2, which illustrates its three main modules.

Module 1 performs source localization and signal estimation by processing data from a distributed microphone array. *Module 2* takes the estimated source locations and signals to characterize the associated VSs. This involves computing the filters $\mathbf{H}_{n,i}(t,\omega,\boldsymbol{r}_q)$ for each source and propagating the sound field to the known microphone positions. This propagation yields an estimate of $\hat{X}_{n,\text{early}}(t,\omega,\boldsymbol{r}_q)$. Thus, an estimate of the diffuse field, $\hat{X}_{\text{diff}}(t,\omega,\boldsymbol{r}_q)$, is obtained as the residual of the obtained propagated signals. The *synthesis* module uses these components to reconstruct the sound field at the VM. It explicitly separates direct and reflected paths for each source and estimates the diffuse field parameters based on the known signals.

An important feature of this framework is its modularity: for example, in environments with microphone-equipped sound sources, where the source locations and signals are already known, Module 1 can be skipped entirely. In such cases, starting directly from Module 2 not only simplifies the pipeline but also improves the overall performance of the algorithm by eliminating potential errors from the estimation stage. As shown in (11), to synthesize the signal at each VM, we need to estimate both the early $S_{n,\text{early}}(t,\omega,\check{\boldsymbol{r}}_v)$ and diffuse $S_{\text{diff}}(t,\omega,\check{\boldsymbol{r}}_v)$ components. To achieve this, we propose the following processing pipeline, which will be discussed in detail in the next sections.

2.2.1 VM components estimation

The model of the direct component $S_{n,\text{direct}}(t,\omega,\check{\boldsymbol{r}}_v)$ is described in (5). The parameters characterizing the direct sound component of a VM are the source location $\boldsymbol{r}'_n$ and the exterior sound field coefficients $\beta^n_{\ell,\mu}(t,\omega)$. The positions $\boldsymbol{r}'_n$ of the sources can be estimated using many different algorithms in the literature, such as [45, 46, 48, 60–62]. However, in this work, we simplify the discussion by considering the source locations as known.

The estimation of the exterior sound field coefficients $\beta^n_{\ell,\mu}(t,\omega)$ from the microphone signals requires the knowledge of the direct sound component $X_{n,\text{dir}}(t,\omega,\boldsymbol{r}_q)$

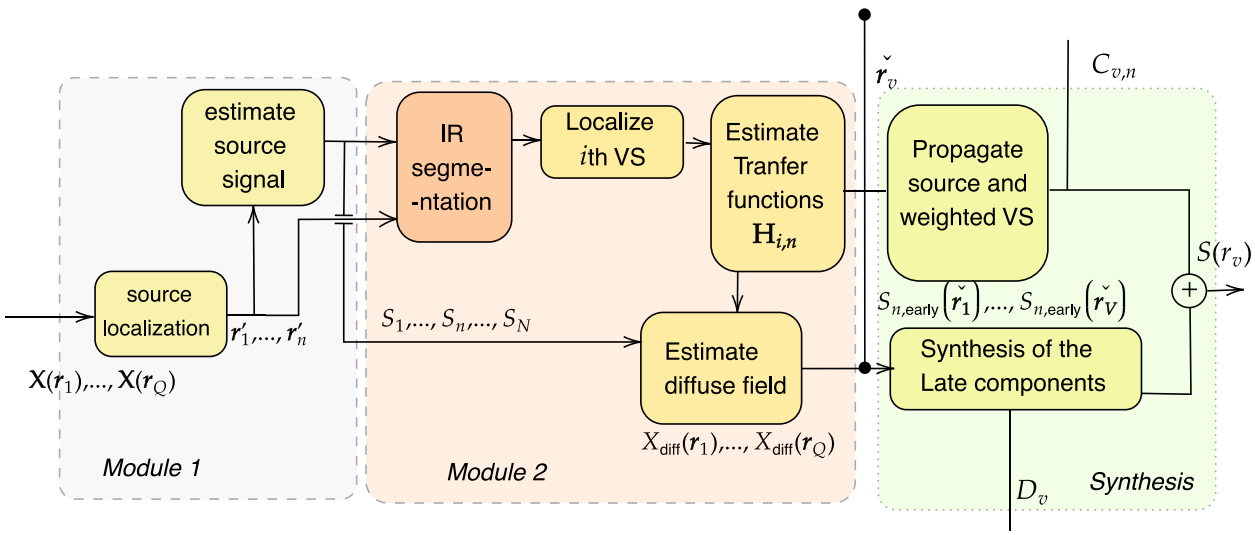

Fig. 2 Block diagram of the proposed modular ParaDER framework for sound field reconstruction. The process encompasses three main stages: source localization and signal estimation (*Module 1*), characterization of virtual sources and diffuse field (*Module 2*), and *synthesis* of the sound field at the virtual microphone location. A detailed description of the signal flow and interactions between modules is provided in Sect. 2.2

at each microphone (see (10)). However, only the microphone signals $X(t,\omega,\boldsymbol{r}_q)$ are directly available. It follows that a procedure for estimating the direct and the diffuse components from $X(t,\omega,\boldsymbol{r}_q)$ is required.

The algorithm for the estimation of the direct components is described in Sec. 3.1.

Once the model of the direct component is estimated, we estimate $H_{i,n}(t,\omega,\check{\boldsymbol{r}}_v)$ from the known signals $X(t,\omega,\boldsymbol{r}_q)$ to reconstruct the ER component on the vth VM. The procedure is discussed in detail in Sec. 3.3.

The diffuse sound component $S_{\text{diff}}(t,\omega,\check{\boldsymbol{r}}_v)$, can be estimated from the microphone diffuse sound components $X_{\text{diff}}(t,\omega,\boldsymbol{r}_q)$, as detailed in Sec. 4.3. Inputs for the estimation of the diffuse components, as shown in Fig. 2, are the VM position $\check{\boldsymbol{r}}_v$, the microphone positions $\boldsymbol{r}_q$ and the VM sensitivity to the diffuse field $D_v(\omega)$.

3 Parameter estimation

In this section, we describe the first phase of the proposed virtual miking approach, which consists of estimating the model parameters introduced in Sect. 2.1. A crucial prerequisite is the knowledge of the physical source locations, since these positions determine the spherical harmonic decomposition used for both direct and ER modeling. In this work, we assume the source locations are known in order to better highlight the contribution of the ParaDER framework itself. Nonetheless, this assumption does not limit the general applicability of the method: the framework can seamlessly incorporate any existing source localization technique to provide the required inputs.

In particular, several methods from the literature can be applied depending on the scenario and available data: geometry-based localization approaches such as the Distributed Ray Space Transform (DRST) [18, 19, 63] and Steered Response Power (SRP) [64, 65]; data-driven approaches such as deep learning methods [66–69]; and hybrid strategies combining beamforming and sparse recovery [48, 60, 61]. These techniques can serve as a front-end to ParaDER, supplying source position estimates that feed into the subsequent parameter estimation stages. By assuming known source positions in our experiments, we isolate and evaluate the performance of ParaDER itself, while maintaining compatibility with state-of-the-art localization strategies. Thus, as underlined in Sect. 2.2, the parameters that have to be estimated are the coefficients of the sources (Sect. 3.1), the direct, the ER (Sect. 3.3), and the diffuse components at each microphone.

3.1 Estimation of early component

Assuming the sources locations to be known, we address the problem of estimating the direct component of a microphone signal, namely $X_{n,\text{dir}}(t,\omega,\boldsymbol{r}_q)$ from the recorded microphone signal $X(t,\omega,\boldsymbol{r}_q)$. This is a crucial step of the process, as the knowledge of $X_{n,\text{dir}}(t,\omega,\boldsymbol{r}_q)$ is required to estimate the exterior sound field coefficients of the sources and relative VSs for the estimation of the VM early $S_{\text{early}}(t,\omega,\check{\boldsymbol{r}}_v)$.

As discussed in [44], the direct component can be estimated using spectral subtraction or Wiener filtering [70–72]. We briefly summarize the approach presented in [44] to highlight key technical aspects that will also be relevant for ER extraction. It is worth noting that this method attenuates the contribution of the diffuse component but does not yield a purely direct component. Therefore, in this paper, we employ it to obtain a first approximation of the $X_{n,\text{early}}(t,\omega,\boldsymbol{r}_q)$ sound field component. Following [50] and [73] we can obtain an estimate of the direct sound component $X_{n,\text{early}}(t,\omega,\boldsymbol{r}_q)$ at the position $\boldsymbol{r}_q$ as the output of a squared root Wiener filter whose coefficients are computed as [50, 74]

$$G_{\text{early}}(t,\omega,\boldsymbol{r}_q)=\sqrt{1-\frac{1}{\text{CDR}(\text{t},\omega,\boldsymbol{r}_\text{q})+1}}, \tag{12}$$

where the Coherence to Diffuse Ratio $\text{CDR}(\text{t},\omega,\boldsymbol{r}_\text{q})$ is the time-frequency dependent signal to diffuse ratio at the qth microphone, defined as

$$\text{CDR}(\text{t},\omega,\boldsymbol{r}_\text{q})=\frac{\Phi_{\text{early,qq}}(\text{t},\omega)}{\Phi_{\text{diff,qq}}(\text{t},\omega)}. \tag{13}$$

Here $\Phi_{\text{early,qq}}$ and $\Phi_{\text{diff,qq}}$ are the auto-power spectra of the direct and diffuse components, respectively, and are defined as

$$\begin{aligned}\Phi_{\text{early,qq}}(t,\omega)&=E\{X_{n,\text{early}}(t,\omega,\boldsymbol{r}_q)X^{\star}_{n,\text{early}}(t,\omega,\boldsymbol{r}_q)\}\\ \Phi_{\text{diff,qq}}(t,\omega)&=E\{X_{\text{diff}}(t,\omega,\boldsymbol{r}_q)X^{\star}_{\text{diff}}(t,\omega,\boldsymbol{r}_q)\}.\end{aligned} \tag{14}$$

As demonstrated in [50], $\text{CDR}(\text{t},\omega,\boldsymbol{r}_\text{q})$ can be estimated using the coherence functions of the microphone signal and diffuse noise

$$\begin{aligned}\text{CDR}(\boldsymbol{r}_\text{q})=&\frac{\Gamma_{\text{diff,ql}}\Re\{\widehat{\Gamma}_{\text{ql}}\}-|\widehat{\Gamma}_{\text{ql}}|^2}{|\widehat{\Gamma}_{\text{ql}}|^2-1}-\\ &-\frac{\sqrt{(\Gamma_{\text{diff,ql}}\Re\{\widehat{\Gamma}_{ql}\})^2-(\Gamma_{\text{diff,ql}}|\widehat{\Gamma}_{ql}|)^2+\Gamma^2_{\text{diff,ql}}-2\Gamma_{\text{diff,ql}}\Re\{\widehat{\Gamma}_{ql}\}+|\widehat{\Gamma}_{ql}|^2}}{|\widehat{\Gamma}_{ql}|^2-1}\end{aligned} \tag{15}$$

where we dropped the arguments t,ω for readability and

$$\hat{\Gamma}_{ql}(t,\omega)=\frac{\Phi_{ql}(t,\omega)}{\sqrt{\Phi_{qq}-\Phi_{N,qq}(t,\omega)}\sqrt{\Phi_{ll}(t,\omega)-\Phi_{N,ll}(t,\omega)}}, \tag{16}$$

$$\Phi_{ql}=E\{X(t,\omega,\boldsymbol{r}_q),X^*(t,\omega,\boldsymbol{r}_l)\} \tag{17}$$

with $\Phi_{N,qq}(t,\omega)$ and $\Phi_{N,ll}(t,\omega)$ defined as in (2). Under the assumption of a spherically isotropic sound field, as

in (1), the diffuse noise coherence function $\Gamma_{\text{diff,ql}(\omega)}$ can be modeled following [50]

$$\Gamma_{\text{diff,ql}}(\omega) = \frac{\Phi_{\text{diff,ql}}(t,\omega)}{\sqrt{\Phi_{\text{diff,qq}}(t,\omega)\Phi_{\text{diff,ll}}(t,\omega)}} = \frac{\sin\left(kd_{ql}\right)}{kd_{ql}}, \tag{18}$$

where

$$\begin{aligned}\Phi_{\text{diff,ql}}(t,\omega) &= E\{X_{\text{diff}}(t,\omega,\boldsymbol{r}_q)X^{\star}_{\text{diff}}(t,\omega,\boldsymbol{r}_l)\},\\ d_{ql} &= \|\boldsymbol{r}_q - \boldsymbol{r}_l\|_2\end{aligned} \tag{19}$$

with $\|\cdot\|_2$ the L-2 norm of a vector. In order to determine the activity or inactivity of the sources, we use the voice activity detector [75].

Once CDR is estimated at the qth microphone, the Wiener filter coefficients can be computed using (12) to extract its direct component. A more practical implementation, as noted in [50], is given by [74]

$$G_{\text{early}}(t,\omega,\boldsymbol{r}_q) = \max\left\{\text{G}_{\text{min}}, 1 - \sqrt{\frac{\mu}{\text{CDR(t},\omega,\boldsymbol{r}_\text{q})+1}}\right\}, \tag{20}$$

where μ controls noise subtraction and G_{min} sets a lower bound to reduce artifacts.

Finally, the filter in (20) is used to compute the direct signal component at the qth microphone through [50]

$$\hat{X}_{\text{early}}(t,\omega,\boldsymbol{r}_q) = G_{\text{early}}(t,\omega,\boldsymbol{r}_q)U(t,\omega,\boldsymbol{r}_q), \tag{21}$$

considering q and l as microphones belonging to the same array,

$$U(t,\omega,\boldsymbol{r}_q) = \sqrt{\frac{Z\left(t,\omega,\boldsymbol{r}_q\right) + Z(t,\omega,\boldsymbol{r}_l)}{2}}e^{j\arg\{\text{X(t},\omega,\boldsymbol{r}_\text{q})\}}, \tag{22}$$

with $Z\left(t,\omega,\boldsymbol{r}_q\right) = |X\left(t,\omega,\boldsymbol{r}_q\right)|^2 - \hat{\Phi}_{N,qq}(t,\omega)$, $Z(t,\omega,\boldsymbol{r}_l) = |X(t,\omega,\boldsymbol{r}_l)|^2 - \hat{\Phi}_{N,ll}(t,\omega)$ and $\arg\{\cdot\}$ the operator that takes the argument of a complex number. The spatial magnitude averaging performed in (22) is typically used in order to reduce the variance of the estimates for microphone array post-filters [76, 77].

It is important to note that while this filtering approach attenuates the contribution of the diffuse field for the microphone pair ql, it does not effectively isolate the direct component, as ER are included in $\hat{X}_{\text{early}}$. Indeed, from the perspective of the qth microphone capsule, ERs can be modeled to be delayed versions of the source signal and, as such, do not satisfy the coherence condition defined in (18). This assumption is verified with very good approximation in environments with planar walls, floor, and ceiling.

Therefore, an additional step is required to extract the direct signal from $\hat{X}_{\text{early}}$. We do so by estimating the exterior sound field coefficients for each source in the scene.

3.2 Source sound field coefficients estimation

Let us introduce the vector $\hat{\mathbf{x}}_{\text{early}}(t,\omega)$, which contains the direct component estimates for all microphones (21). The direct sound components acquired by the microphones are given by

$$\begin{aligned}\hat{\mathbf{x}}_{\text{early}}(t,\omega) &= \left[\widehat{\mathbf{Y}}_1(\omega)\widehat{\mathbf{Y}}_2(\omega)\cdots\widehat{\mathbf{Y}}_N(\omega)\right]\begin{bmatrix}\boldsymbol{\beta}_1(t,\omega)\\ \vdots\\ \boldsymbol{\beta}_N(t,\omega)\end{bmatrix}\\ &= \widehat{\mathbf{Y}}(\omega)\boldsymbol{B}(t,\omega).,\end{aligned} \tag{23}$$

where

$$\boldsymbol{\beta}_n(t,\omega) = \left[\beta^n_{00}(t,\omega), \beta^n_{0-1}(t,\omega),\ldots,\beta^n_{LL}(t,\omega)\right]^{\top} \tag{24}$$

is the vector of the coefficients of the spherical harmonic for the nth source, and $\widehat{\mathbf{Y}}_n(\omega)$ is the matrix containing the spherical harmonics related to the nth source given its location $\boldsymbol{r}'_n$ Since $\widehat{\mathbf{Y}}(\omega)$ is known from source locations, an estimate

$$\hat{\boldsymbol{B}}(t,\omega) = \arg\min||\hat{\mathbf{Y}}\boldsymbol{B}(t,\omega) - \hat{\mathbf{x}}_{\text{early}}|| \tag{25}$$

can be addressed using an optimization method as demonstrated in [44].

Since this method optimizes along the source directions, it not only retrieves the exterior field coefficients of the source but also suppresses the contribution of ER in the estimated direct component $\hat{X}_{\text{early}}(t,\omega,\boldsymbol{r}_q)$.

3.3 Estimation of early reflections component

In (1), we defined as ER component as the part of the microphone signal that is correlated with the source signal. One approach to estimating this correlation is to recover an estimate of the RIR. By definition, the peaks in the RIR numerically represent the correlation between the source signal and the recorded signal.

However, estimating the RIR requires knowledge of the source signal modeled as a point source. This is necessary to preserve the spatial relationship with the VS and to avoid distortion in the correlation estimation caused by the directional patterns of sources. Specifically, these point-source-like signals can be approximated using the zero-order spherical harmonic coefficients, which represent the omnidirectional component of the sound field. This component reasonably approximates a point source. Following the approach described in [37], we estimate the point source-like signal as

$$\tilde{S}_n(t,\omega) = \hat{\beta}^n_{00}/c_{00}, \tag{26}$$

where $c_{00} = -ik$. Thus, $\tilde{S}_n(t,\omega)$ is used to deconvolve the observed signal $X(t,\omega,\boldsymbol{r}_q)$ obtaining an estimate of the

RIR. A widely used method for this purpose is the function $\mathcal{G}(\cdot)$ derived from the Generalized Cross-Correlation [78].

$$\mathcal{R}_{q,n}(\tau) = \mathcal{G}(S_n(t,\omega), X(t,\omega,\boldsymbol{r}_q)), \tag{27}$$

where $\mathcal{R}_{q,n}(\tau)$ is the estimate of the impulse response in the time domain τ for the qth capsule with respect to the nth sound source.

Therefore, we define

$$\boldsymbol{\mathcal{R}}_{a,n}(\tau) = [\mathcal{R}_{q,n}(\tau)]_{q\in\mathcal{N}_a}, \tag{28}$$

aggregating the RIRs corresponding to all microphone capsules q into their respective microphone arrays a, specifically in the set $\mathcal{N}_a$. Given $\boldsymbol{\mathcal{R}}_{a,n}(\tau)$ we perform peak detection over a threshold to extract the part of the signal that is more correlated to the source signal. We consider the part of the $\boldsymbol{\mathcal{R}}_{a,n}(\tau)$ for $\tau > |\boldsymbol{r}'_n - \boldsymbol{r}_q|/c$, since it is related to the direct path. Thus, the number of virtual sources, $\hat{I}_{a,n}$, is determined using a peak energy threshold of -15dB compared to the first peak energy, obtaining $\hat{I}_{a,n}$ peaks related delay values $\delta_{i,a,n}$ from $i = 1,\dots,\hat{I}_{a,n}$. Each peak is then isolated by segmenting $\boldsymbol{\mathcal{R}}_{a,n}$ with a Hann window $W_{i,a,n}$ of length W (samples), centered at $\delta_{i,a,n}$. The window length W is a fixed parameter chosen to be long enough to capture the energy of a single reflection yet short enough to temporally resolve distinct early reflections; a value of $W = 128$ samples (≈ 2.7ms at 48 kHz) was used throughout this work. This process yields $\hat{I}_{a,n}$ segmented ER.

For each segment, we perform the DOA estimation [79], resulting in the estimate of $\theta_{i,a,n}, \varphi_{i,a,n}$ azimuth and elevation, indeed, $\rho_{i,a,n} = \delta_{i,a,n}c$. Thus, for each array a we obtain $I_{a,n}$ estimates of VS locations $\boldsymbol{r}''_{i,a,n} = [\theta_{i,a,n}, \varphi_{i,a,n}, \rho_{i,a,n}]$. We summarize this operation in the function

$$\mathcal{V}(\boldsymbol{\mathcal{R}}_{a,n} W_{i,a,n}) = \boldsymbol{r}''_{i,a,n}. \tag{29}$$

Since the localization of the VSs is affected by noise, to improve the accuracy we perform K-Nearest Neighbors (KNN) [80] aggregating in I_n clusters the found VSs

$$\mathcal{N}_k^{\epsilon}(\boldsymbol{r}''_{i,n}) = \left\{ \boldsymbol{r}''_{i,a,n} \in \mathcal{P} \setminus \{\boldsymbol{r}''_{i,n}\} \;\middle|\; |\boldsymbol{r}''_{i,n}, \boldsymbol{r}''_{i,a,n}|^2 \le \epsilon \right\}.$$

The set $\mathcal{P}$ is the entire collection of all estimated VS positions. For a candidate point $\boldsymbol{r}''_{i,n}$, its ϵ-neighborhood $\mathcal{N}_k^{\epsilon}$ is defined as all other points in $\mathcal{P}$ within a Euclidean distance of ϵ.

Thus, from now on, we will refer to VSs as $\boldsymbol{r}''_{i,n}$ removing the subscript a.

Given the location of each VS, we need to estimate the transfer function $H_{n,i}(t,\omega,\boldsymbol{r})$ that describes the relationship between the nth source and the ith VS in the location $\boldsymbol{r}$. Thus, we define

$$\hat{H}_{n,i}(t,\omega,\boldsymbol{r}_a) = \sqrt{||\hat{X}_{i,n}(t,\omega,\boldsymbol{r}_a)||^2 / ||\hat{X}_{n,\mathrm{dir}}(t,\omega,\boldsymbol{r}_a)||^2}, \tag{30}$$

where $\boldsymbol{r}_a = \sum_{q\in\mathcal{N}_a} \boldsymbol{r}_q/M$ is the location of the center of the array ath, $\hat{X}_{i,n}(t,\omega,\boldsymbol{r}_a) = \mathcal{S}(\mathcal{R}_{a,n} W_{i,a,n}) X(t,\omega,\boldsymbol{r}_a)$, $\hat{X}_{n,\mathrm{dir}}(t,\omega,\boldsymbol{r}_a) = \sum_{\ell=0}^{L}\sum_{\mu=-\ell}^{\ell} \hat{\beta}^n_{\ell\mu}(t,\omega) h_\ell(k\rho_{a,n}) Y_{\ell\mu}(\theta_{a,n},\phi_{a,n})$, with $\hat{\beta}_{\ell,\mu}$ estimated in (25), $X(t,\omega,\boldsymbol{r}_a) = \sum_{q\in\mathcal{N}_a} X(t,\omega,\boldsymbol{r}_q)/M$, and $\mathcal{S}(\cdot)$ is the Short Time Fourier Transform.

3.4 Estimation of diffuse component

The diffuse component of the microphone signal can be obtained using the filter [73]

$$G_{\mathrm{diff}}(t,\omega,\boldsymbol{r}_q) = \sqrt{1 - \left[\hat{G}_{\mathrm{early}}(t,\omega,\boldsymbol{r}_q)\right]^2}, \tag{31}$$

where $\hat{G}_{\mathrm{early}}(t,\omega,\boldsymbol{r}_q)$ is defined in similarity to (20) substituting $\Phi_{\mathrm{early,qq}}$ with $\hat{\Phi}_{\mathrm{early,qq}} = E\{\hat{X}_{n,\mathrm{early}}(t,\omega,\boldsymbol{r}_q), \hat{X}^*_{n,\mathrm{early}}(t,\omega,\boldsymbol{r}_q)\}$. It follows that an estimate of $X_{\mathrm{diff}}(t,\omega,\boldsymbol{r}_q)$ can be obtained as

$$\hat{X}_{\mathrm{diff}}(t,\omega,\boldsymbol{r}_q) = G_{\mathrm{diff}}(t,\omega,\boldsymbol{r}_q) U(t,\omega,\boldsymbol{r}_q), \tag{32}$$

where $U(t,\omega,\boldsymbol{r}_q)$ is defined in (22).

4 Synthesis

4.1 Synthesis of the direct component

An estimate $\hat{S}_{n,\mathrm{dir}}(t,\omega,\check{\boldsymbol{r}}_\nu)$ of the direct sound component at the νth VM due to the nth source can be obtained by exploiting the model in (5). More precisely, given the source locations $\hat{\boldsymbol{r}}'_n$ and the set of exterior field coefficients $\hat{\boldsymbol{\beta}}_n(t,\omega)$ obtained in (23), $\hat{S}_{n,\mathrm{dir}}(t,\omega,\check{\boldsymbol{r}}_\nu)$ is obtained through

$$\hat{S}_{n,\mathrm{dir}}(t,\omega,\check{\boldsymbol{r}}_\nu) = C_\nu(\omega) \sum_{\ell=0}^{L}\sum_{\mu=-\ell}^{\ell} \hat{\beta}^n_{\ell\mu}(t,\omega) h_\ell\left(k\hat{\check{\rho}}_{\nu,n}\right) Y_{\ell\mu}\left(\hat{\check{\theta}}_{\nu,n}, \hat{\check{\phi}}_{\nu,n}\right), \tag{33}$$

where $\hat{\check{\rho}}_{\nu,n}$, $\hat{\check{\theta}}_{\nu,n}$, and $\hat{\check{\phi}}_{\nu,n}$ are the estimates of $\check{\rho}_{\nu,n}$, $\check{\theta}_{\nu,n}$, and $\check{\phi}_{\nu,n}$ in (5) and can be computed by inserting in (3) the estimate $\hat{\boldsymbol{r}}'_n$ of the nth source location.

The term $C_\nu(\omega)$ in (11) models the VMs pick-up pattern. Usually this term can be expressed as a function $f(\cdot)$ that depends on the frequency ω, the angle between the νth VM and the nth source. More details about the pattern model are discussed in [44].

4.2 Synthesis of the early reflections component

As described in (9), to synthesize the ER, we need the set of sources' exterior field coefficients $\hat{\boldsymbol{\beta}}_n(t,\omega)$, the location of the VSs $\boldsymbol{r}''_{i,n}$, and the transfer matrix $H_{i,n}(t,\omega,\check{\boldsymbol{r}}_\nu)$.

Assuming that the sound energy propagation has linear decay in space, we can approximate $H_{n,i}(t,\omega,\check{\boldsymbol{r}}_\nu)$ as

$$\tilde{H}_{n,i}(t,\omega,\check{\boldsymbol{r}}_\nu)=\sum_{a=1}^{A}\varpi_a\hat{H}_{n,i}(t,\omega,\boldsymbol{r}_a), \tag{34}$$

where $\sum_{a=1}^{A}\varpi_a=1$. We choose the weights to be inversely proportional with respect to the distance between the νth VM and the ath array center, i.e.,

$$\varpi_a(\omega)=\frac{1}{\|\boldsymbol{r}_a-\check{\boldsymbol{r}}_\nu\|_2}\left(\sum_{p=1}^{A}\frac{1}{\|\boldsymbol{r}_p-\check{\boldsymbol{r}}_\nu\|_2}\right)^{-1}. \tag{35}$$

Thus, early reflection term related to the nth source on the νth VM location $\check{\boldsymbol{r}}_\nu$ in (9) becomes

$$\hat{S}_{n,\mathrm{ER}}(t,\omega,\check{\boldsymbol{r}}_\nu)=C_\nu(\omega)\sum_{i}^{I_n}\tilde{H}_{n,i}(t,\omega,\check{\boldsymbol{r}}_\nu)\sum_{\ell=0}^{L}\sum_{\mu=-\ell}^{\ell}\beta_{\ell\mu}^{n}(t,\omega)h_\ell\left(k\check{\rho}_{\nu,i,n}\right)Y_{\ell\mu}\left(\check{\theta}_{\nu,n},\check{\phi}_{\nu,n}\right). \tag{36}$$

This equation enables the estimation of the reflected contribution from a specific source n, based on the corresponding estimated parameters. The image source method models reflected sound as originating from a VS. This VS mirrors the original source's characteristics but is modified by the properties of the reflective surface. Consequently, we model each reflection by placing a VS that emits the original signal, filtered to account for the surface's acoustic effects, which then propagates toward the VM's position.

This approach allows us to adopt a simplified propagation model in which the acoustic wave propagation is characterized by the Hankel function h_ℓ wavefronts in free-space conditions. To incorporate non-idealities of the reflective surface, such as frequency-dependent absorption or diffusion due to material properties or geometric irregularities, we introduce the matrix $\tilde{H}_{n,i}$. This matrix models the surface's frequency response, accounting for effects like high-frequency attenuation caused by rough or absorptive surfaces.

4.3 Synthesis of the diffuse component

In order to synthesize the diffuse component $S_{\mathrm{diff}}(t,\omega,\check{\boldsymbol{r}}_\nu)$ of the VM, we use the estimates of the diffuse signal at each microphone obtained in (32). More precisely, given the estimates $\hat{X}_{\mathrm{diff}}(t,\omega,\boldsymbol{r}_q)$, we compute the power of the diffuse signal component in $\check{\boldsymbol{r}}_\nu$ as [15]

$$E\{|S_{n,\mathrm{diff}}(t,\omega,\check{\boldsymbol{r}}_\nu)|^2\}=\sum_{q=1}^{Q}\varpi_q(\omega)E\{|\hat{X}_{n,\mathrm{diff}}(t,\omega,\boldsymbol{r}_q)|^2\}. \tag{37}$$

For what concerns the estimation of the phase of the diffuse signal component, according to [44], plausible results can be achieved by using the phase of the nearest microphone.

Finally, as defined in (11), the function $D_\nu(\omega)$ controls the sensitivity of the VM to the diffuse field. In the most general case, this function can be arbitrarily designed to resemble the characteristics of a binaural or cardioid microphone [19, 81, 82].

5 Validation and results

To comprehensively evaluate the proposed method, we assess its performance through numerical analysis and perceptual investigation. This section presents the two validation approaches and the results.

First, we conduct a simulation campaign within a shoebox room, where a limited number of HOMs measure a controlled area. Second, we perform perceptual experiments using an augmented reality visor and headphones. The simulation setup is detailed in Sect. 5.1, while Sect. 5.2 defines the metrics used to assess the performance of the VM signal reconstruction. The simulation results are discussed in Sect. 5.3. Following this, we provide a deep discussion of the perceptual experiment in Sec. 5.5.

5.1 Simulation setup and parameters

The simulation setup, illustrated in Fig. 3, consists of $A=9$ HOMs, each with a radius of 0.025 m. To estimate a second-order spherical harmonic expansion, at least $M\geq 9$ microphones are required per array. To improve numerical stability, we employ $M=12$ omnidirectional microphones for each HOM. The microphone capsules are arranged according to the geometry proposed in [83], resulting in a total of $Q=A\times M=108$ microphones in the scene.

We simulate a scenario similar to a conference room, where the audience is seated on an inclined surface. Accordingly, the HOMs are also positioned on this incline, as illustrated in Fig. 3. The lowest HOM is positioned at least 1 and 0.5 meters away from the source and the walls, respectively, with its center height located as $z/2-0.5$ m. The highest and farthest HOM is located at $\boldsymbol{r}_A=[d-0.5\,\mathrm{m},\ l-0.5\,\mathrm{m},\ h/2+1\,\mathrm{m}]^T$, where d, l refers to the length and width of the room as in in Fig. 3. To evaluate spatial accuracy, we consider a single source scenario, though experiments involving multiple sources

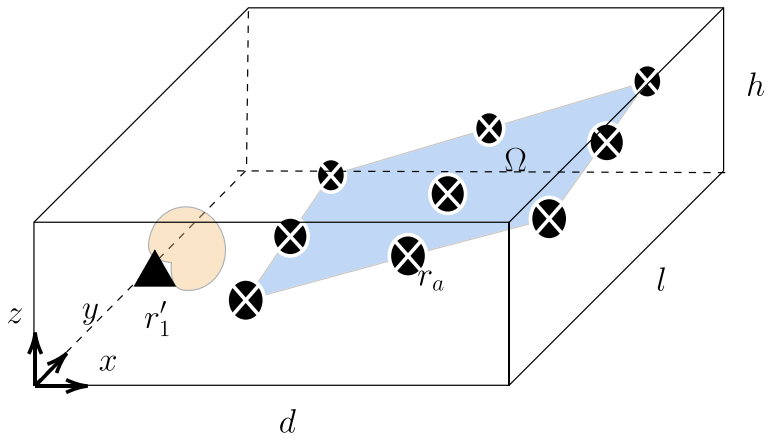

Fig. 3 3D graphical representation of the simulation setup. The source, represented with ▲, is located at $\boldsymbol{r}'_1 = [1.5, 2, h/2]^T$ m, and its directivity pattern is $\pi/4$ w.r.t. x axes oriented, depicted as a red shaded area. The known HOMs centered in $\boldsymbol{r}_a$ and represented as a black circle with a white cross are located on an inclined plane on the control surface Ω inside the room

can be found in [44]. Thus, we employed the same source localization method discussed in [44]. The source is placed at $\boldsymbol{r}'_1 = [1.5, 2, h/2]^T$ m and emits a speech signal from [84]. It has a first-order cardioid directivity with a maximum energy emission direction of 45° azimuth and 0° elevation.

In the Ω region, a dense 7×7 grid of equally spaced virtual HOMs was placed. Each virtual νth HOM consists of $\check{Q}_\nu = 12$ VMs of the same type as the known HOMs, serving to assess the accuracy of the higher-order reconstruction. Thus, a total of $V = 588$ VMs are reconstructed in the room.

In order to test the system in a range of acoustic conditions, the sources and microphones configuration is accommodated in rooms with variable size. More specifically, $\Xi = 100$ different room configurations, with randomly selected dimensions: $4\,\text{m} \le d \le 9\,\text{m}$, $3\,\text{m} \le l \le 9\,\text{m}$, and $3\,\text{m} \le h \le 6\,\text{m}$ have been simulated. The HOM positions were adjusted based on room size, while the source location remained fixed. Such rooms are also characterized by reverberation time $T60$ randomly selected from 0.4 to 1.6 s. The microphone signals of the HOMs in (1) are generated by convolving the source signals with the RIRs, which are computed using the image source method [51] as implemented in [85].

The additive noise term in (1) is modeled as random white Gaussian noise, with its variance adjusted to achieve a signal-to-noise ratio (SNR) of 60 dB.

All signals are sampled at 48 kHz, and their time-frequency representation is obtained via an 8192-point Short-Time Fourier Transform (STFT), using a 512-time bin Hamming window with an 87.5% overlap, applied consistently in both the analysis and synthesis stages. The coherence-based filter parameters in Eq. (20) were set to $\mu = 1.5$ and $G_{\min} = -15$dB. For the RIR segmentation in Sec. 3, a Hann window of fixed length $W = 128$ samples was used to isolate ERs, which were then clustered with a distance threshold of $\epsilon = 0.1$m.

5.2 Metrics

To assess the accuracy of the reconstructed sound field, we employ the following key metrics, detailed in the next subsections:

- Power spectral density (PSD) Distribution, to evaluate the spatial distribution of energy in the frequency domain;
- Diffuseness[86], to quantify the proportion of non-directional energy in the reconstructed field;
- Early decay time (EDT)[87], derived from the impulse response obtained via deconvolution of the source signal and the estimated signal, characterizing the temporal decay properties of the sound field.

To evaluate the metrics above, the SH decomposition of the reconstructed HOMs is needed. It is expressed as

$$\boldsymbol{\alpha}(t, \omega, \check{\boldsymbol{r}}_\nu) = [\alpha_{00}(t, \omega, \check{\boldsymbol{r}}_\nu), \alpha_{-11}(t, \omega, \check{\boldsymbol{r}}_\nu), \ldots, \alpha_{LL}(t, \omega, \check{\boldsymbol{r}}_\nu)]^\top, \tag{38}$$

where $\check{\boldsymbol{r}}_\nu$ represents the centroid of the νth virtual HOM, computed as the mean position of its constituent VMs

$$\check{\boldsymbol{r}}_\nu = \frac{1}{M} \sum_{\nu=1}^{12} \check{\boldsymbol{r}}_\nu. \tag{39}$$

Following the formulation in [88], the SHC $\alpha_{nm}(t,\omega,\check{\boldsymbol{r}}_\nu)$ are given by

$$\alpha_{nm}(t,\omega,\check{\boldsymbol{r}}_\nu) = \frac{1}{b_n(\omega\check{\boldsymbol{r}}_\nu)} \sum_{\nu=1}^{Q_\nu} \hat{S}(t,\omega,\check{\boldsymbol{r}}_\nu) Y_{nm}^*(\bar{\theta}_\nu,\bar{\phi}_\nu) w_\nu. \tag{40}$$

Here, $b_n(\omega r_\nu)$ denotes the spherical Bessel function, w_c is a microphone-specific weight ensuring numerical consistency, $(\bar{\theta}_\nu,\bar{\phi}_\nu)$ are the azimuth and elevation angles of the relative position vector $\bar{\boldsymbol{r}}_\nu = \boldsymbol{r}_\nu - \boldsymbol{r}_\nu$.

5.2.1 PSD related metrics

We propose two different metrics related to the PSD. One is obtained from a comparison between the ground truth PSD and the estimated PSD, another is obtained by masking the source direction to emphasize the non-directional component of the sound field. In particular, the PSD of a multichannel acquisition is defined accordingly to MUSIC algorithm in SHD [79] considering the frequency band from 100 to 4000 Hz as

$$\Psi_\nu(\theta,\phi) = \sum_\omega \frac{1}{\mathbf{Y}^\top(\theta,\phi)\mathbf{V}_{n,\nu}\mathbf{V}_{n,\nu}^\top\mathbf{Y}(\theta,\phi)}, \tag{41}$$

where we dropped the frequency ω argument for readability, $\mathbf{Y}(\theta,\phi) = [Y_{00}(\theta,\phi),\ldots,Y_{LL}(\theta,\phi)]^\top$ represents the real-valued spherical harmonics evaluated at direction (θ,ϕ). $\mathbf{V}_{n,\nu}$ is the noise subspace, obtained from the eigenvalue decomposition of the covariance matrix $\mathbf{C}_\nu = \sum_\omega \sum_t \boldsymbol{\alpha}^\top(t,\boldsymbol{r}_\nu)\boldsymbol{\alpha}(t,\boldsymbol{r}_\nu) \in \mathbb{C}^{(L+1)^2\times(L+1)^2}$, which can also be written as

$$\mathbf{C}_\nu = \mathbf{V}_\nu \Lambda \mathbf{V}_\nu^\top,$$

where Λ is a diagonal matrix containing the eigenvalues of $\mathbf{C}_\nu$. The matrix $\mathbf{V}_\nu$ contains the sorted eigenvectors, and the noise subspace is formed by selecting the eigenvectors corresponding to the lowest eigenvalues:

$$\mathbf{V}_{n,\nu} = [\mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_N],$$

Along with the comparison between regular PSDs, we also compare masked PSDs, where the contribution of the source is suppressed. We do so through a Von Mises distribution, which is fitted to model the dominant peak in the spatial power spectral density (PSD). This distribution acts as a directional analogue of a Gaussian, capturing the concentration of energy around the main peak. The model is defined as

$$\mathcal{V}(\boldsymbol{r}) = \frac{\kappa}{2\pi(e^\kappa - e^{-\kappa})} e^{\kappa \boldsymbol{r}^\top \boldsymbol{r}_1'}$$

where $\boldsymbol{r}_1'$ is the direction corresponding to the peak of the PSD, and κ is the concentration parameter controlling the sharpness of the distribution, i.e., the concentration parameter. After detecting a peak at $\boldsymbol{r}_1'$, the PSD is suppressed at that location using the inverse of the Von-Mises distribution

$$\Psi_{2,\nu}(\theta,\phi) = \frac{\Psi_\nu(\theta,\phi)}{\mathcal{V}(\mathbf{r}) + \xi}$$

where ξ is a small regularization constant. In practice, a grid of (θ,ϕ) is defined with steps of 2[deg] between $-180, 180$[deg] and $-90, 90$[deg] respectively. Thus we obtain $\Psi_\nu, \Psi_{2,\nu} \in \mathbb{R}^{180\times 90}$, describing the sound field power around the νth virtual HOM.

5.2.2 Diffuseness

We use the COMEDIE diffuseness estimator from SH signals described in [86]. This estimator assesses how uniformly the sound energy is distributed across different directions. For each spherical harmonic order ℓ, we extract the corresponding submatrix

$$\mathbf{C}_{\ell,\nu} = \mathbf{C}_\nu(1:(\ell+1)^2, 1:(\ell+1)^2)$$

The spread factor at order ℓ is given by

$$g_\ell = \frac{1}{\lambda_\ell} \sum_{\ell'=1}^{(\ell+1)^2} |\lambda_{\ell'} - \frac{1}{(\ell+1)^2} \sum_{k'=1}^{(\ell+1)^2} \lambda_{k'}|, \tag{42}$$

where λ_is are the eigenvalues of $\mathbf{C}_\ell$ and we dropped the ν virtual HOM index, since it is evaluated for each virtual HOM. Thus, the diffuseness at order ℓ for the virtual HOM ν is

$$D_{\ell,\nu} = 1 - \frac{g_{\ell,\nu}}{2((\ell+1)^2 - 1)}. \tag{43}$$

A higher diffuseness value indicates a more uniform, non-directional sound field, while a lower value suggests the presence of distinct, directional sound sources. This quantification aligns with the perceptual experience of sound: a highly diffuse field envelops the listener, making it challenging to pinpoint the origin of sounds, whereas a less diffuse field perceptually allows for easier localization of sound sources [6].

5.2.3 Early decay time (EDT)

EDT is the time required for the sound energy to decay by 10dB after the initial impulse, typically extrapolated to 60dB for comparison with $T60$. It is a valuable metric for characterizing an RIR, as it better reflects the perceived reverberation in the ER, which significantly impacts clarity and intelligibility in acoustic environments.

To evaluate EDT, the RIR is obtained by deconvolving the output signal of the algorithm with the input source signal. Then the procedure described in [87] is applied.

5.2.4 Normalized mean square error (NMSE)

For each metric (Ψ_ν, $\Psi_{2,\nu}$, $D_{\ell,\nu}$, and EDT), we evaluate the Normalized Mean Squared Error (NMSE). Denoting a generic metric as A, the NMSE is computed as

$$\mathrm{NMSE(A)} = \frac{\|\hat{\mathrm{A}} - \mathrm{A_{GT}}\|^2}{\|\mathrm{A_{GT}}\|^2} \tag{44}$$

where $\hat{A}$ represents the estimated metric obtained from the reconstructed data, and A_{GT} is its corresponding ground truth value. The NMSE provides a normalized measure of error, ensuring comparability across different metrics.

5.3 Simulation results

We compared the proposed algorithm with a similar parametric method presented in [44]. Notably, the method in [44] considers only the direct and diffuse components, without accounting for ER. To distinguish between the two approaches, we refer to the method in [44] as D-D (Direct and Diffuse), while the proposed method parametric modeling of direct sound, early reflection and reverberation, which incorporates ER, is denoted as ParaDER.

In Fig. 4, the results obtained in our simulation campaign are depicted, showing how the NMSE varies in space. In particular, the averaged NMSE is depicted, which is defined as

$$\widehat{\mathrm{NMSE}}(A) = \sum_{\xi=1}^{\Xi} \mathrm{NMSE}(\mathrm{A}_\xi)/\mathrm{E} \tag{45}$$

where ξ is the simulation index, Ξ is the total number of simulations, and A_ξ is the generic metric evaluated in ξth simulation. In the first and second columns of Fig. 4, the spatial distributions of Ψ and Ψ_2 are depicted. Notably, in the region near the source, the PSD of the D-D method is lower than that of the proposed approach. However, in the remaining space, the NMSE decreases. We can interpret this result in terms of how D-D is designed, i.e., a sound field consisting only of direct sound and an isotropic diffuse component, a condition well approximated near the source. In reality, the sound field contains ERs, which can be better captured by masking the source component in the PSD.

Examining the second row, Ψ_2 reveals that the proposed method significantly improves the reconstruction of reflections, enhancing the NMSE on synthesis by more than 10 db.

In the third column, the proposed method surpasses D-D in terms of diffuseness accuracy, indicating that incorporating the ER model in ParaDER leads to a more perceptually accurate spatial reconstruction. Since the results obtained from the diffuseness of orders 1 and 2, i.e., $D_{1,\nu}$ and $D_{2,\nu}$, are similar to the zero order, we depicted here only the zero-order one. A similar improvement is observed for the EDT, confirming that explicitly modeling ER not only enhances objective reconstruction metrics, but also contributes to a more realistic perception of the acoustic scene.

In Fig. 5, the NMSE metrics are shown for various randomly sampled $T60$ values across 100 simulations. For

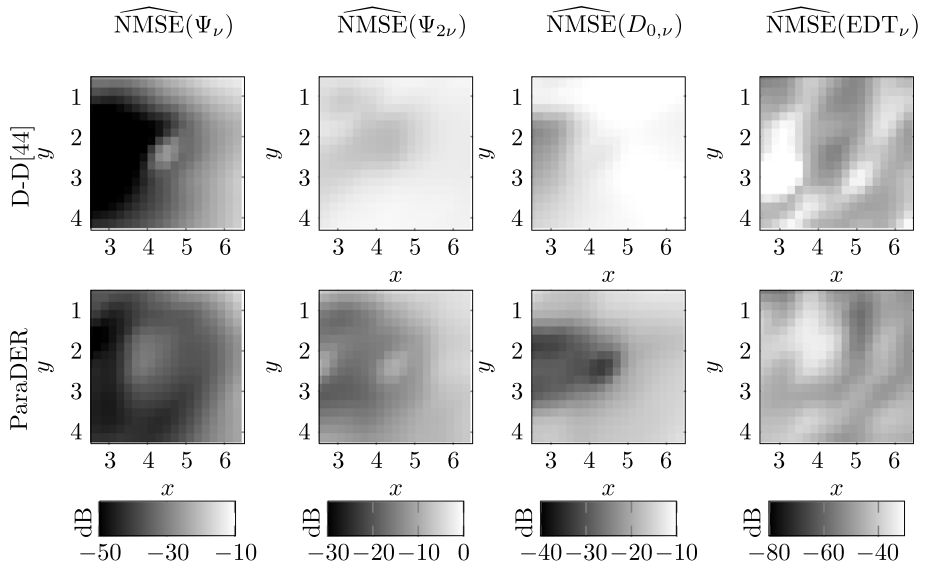

Fig. 4 Top view of the control surface Ω, where the NMSE is averaged over 100 simulations. From left to right, the plots represent the NMSE of the PSD, the NMSE of the PSD with source masking, the NMSE of the diffuseness, and the NMSE of the EDT. The top row corresponds to the D-D method from [44], while the bottom row presents results from the proposed ParaDER method

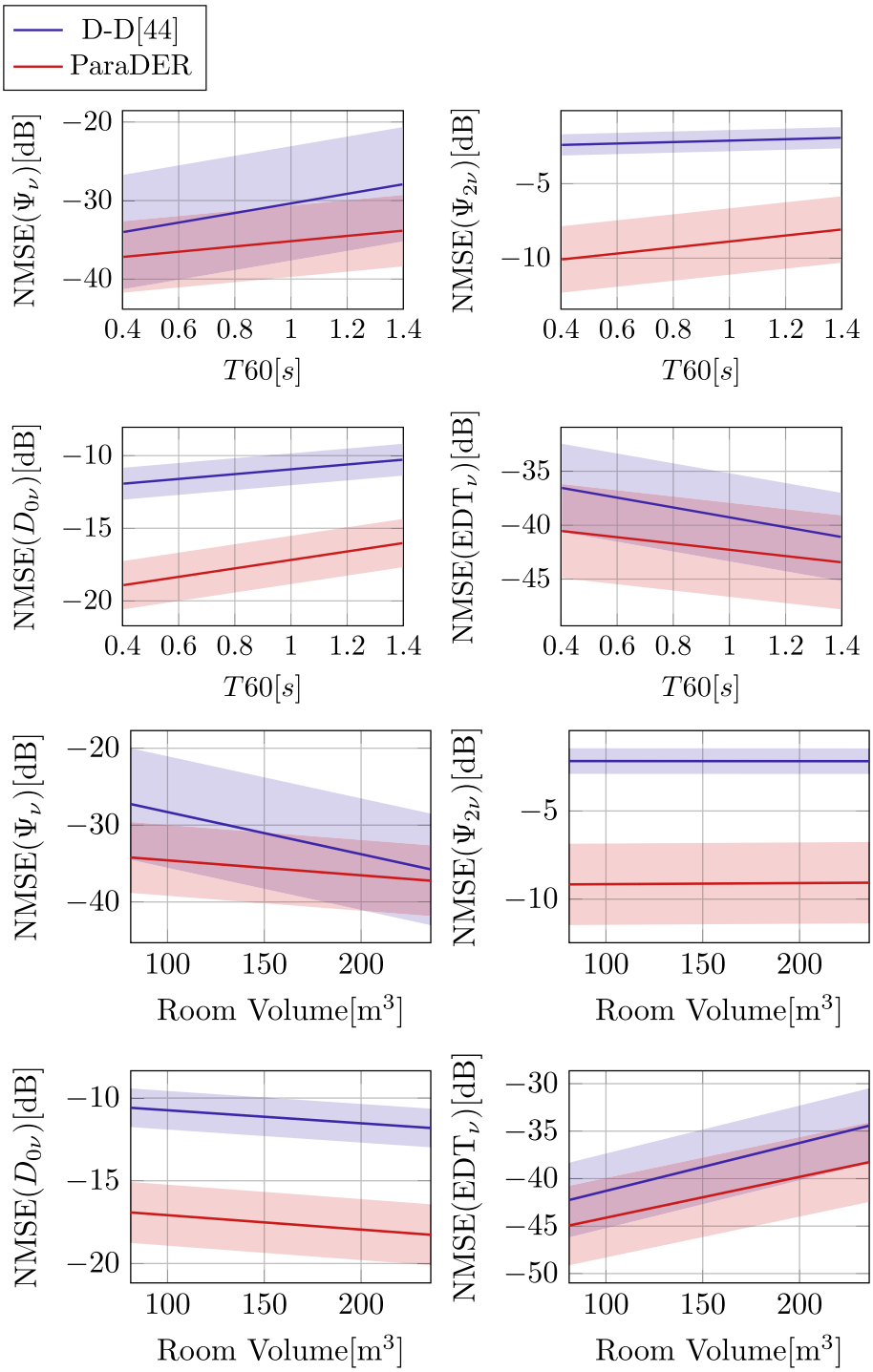

Fig. 5 NMSEs for the considered metrics as a function of $T60$ and the volume of the room. The solid lines depict the first-order polynomial approximation of the NMSEs computed over 100 × 49 simulations (i.e., simulations per number of virtual HOMs). The shaded areas indicate the corresponding standard deviations

each simulation, the 49 HOM metrics are used to compute the mean and standard deviation of the NMSEs.

From Fig. 5, it is evident that for all metrics except EDT, the NMSE increases as $T60$ increases for both models. This trend indicates that higher $T60$ values correspond to more complex acoustic scenarios, where the assumptions of the parametric model become less valid. Nevertheless, the proposed method consistently improves the D-D

method, demonstrating that incorporating ERs enhances the reconstruction quality. Moreover, we found that the NMSE of the EDT decreases as the T60 increases. This is because, in highly reverberant spaces, the sound field is dominated by the diffuse reverberant tail. This diffuse component is more predictable and uniform than individual ER, making it easier for a reconstruction model to accurately estimate the overall decay, including the initial EDT slope.

In the bottom part of Fig. 5, the NMSEs as a function of room volume are presented. We observe that as the room volume increases, both methods show improved accuracy in terms of PSD and diffuseness $D_{0\nu}$. However, the accuracy of the EDT estimation decreases. This is because, in larger spaces, virtual sources are positioned farther from the HOMs, making it more challenging to accurately identify reflections. Additionally, in larger rooms, the diffuse field between distant HOMs becomes more uncorrelated compared to the more coherent diffuse field observed in smaller spaces. This effect is partly due to the higher spatial density of HOMs in smaller environments; since the same number of arrays is distributed over a more compact volume, the measurement grid becomes denser, leading to a more coherent sampling of the diffuse field. These observations highlight the importance of modeling both ER and the diffuse field. Depending on the room's characteristics, an accurate representation of the sound field often requires a weighted combination of these two components.

5.4 Robustness to additive noise

To evaluate the robustness of the proposed method in the presence of additive noise, an extensive Monte Carlo simulation campaign was undertaken. The experimental setup mirrors that described in Sect. 5.1, utilizing a shoebox room with dimensions 6m × 5m × 4m and a reverberation time of $T60 = 0.8$ s. To systematically assess performance degradation, uncorrelated white Gaussian noise was introduced to the simulated microphone signals to achieve a range of signal-to-noise ratio (SNR) conditions from 5 to 60 dB. For each SNR level, the algorithm's performance was quantified by reconstructing the virtual HOMs signal within the target region Ω (illustrated in Fig. 3) and computing the NMSE for the proposed metrics.

The results in Fig. 6 depict the statistical distribution of the NMSE across all tested SNR values. As anticipated, the estimation fidelity is inversely correlated with the noise power. A critical observation is the algorithm's stable performance regime for SNR>20 dB, where the NMSE plateaus. This indicates a degree of inherent noise resilience, suggesting that the proposed method remains effective under realistic acoustic conditions where moderate noise levels are expected. The performance degradation below this threshold follows an expected trend, characterized by a monotonic increase in both the mean and variance of the reconstruction error.

5.5 Real case scenario results

To evaluate the proposed algorithm in a realistic scenario, we conducted a listening experiment using a MUSHRA (Multiple Stimuli with Hidden Reference and Anchor) like test. MUSHRA is a widely used subjective evaluation methodology designed for high-quality audio assessment, where participants rate different stimuli compared to a hidden reference on a continuous scale from 0 to 100.

Our experiment is based on the dataset of a large conference room presented in [49], from which the reference RIRs of the HOMs were obtained for two sound sources. Specifically, we considered $A = 25$ HOMs giving a total of $Q = 25 \times 8$ capsules. To evaluate the reconstruction accuracy, we assume that only 50% of the microphones are available for analyzing the acoustic field, while the remaining 50% are used to provide a ground truth for comparison. Specifically, referring to Fig. 8, two configurations are considered: (i) HOMs with odd indices are used for analysis, and those with even indices for validation, and (ii) the roles are reversed. To generate the test stimuli, the RIRs in the training set were convolved with three different audio signals: male and female anechoic speech signals from [89] and a music anechoic signal from [90]. The resulting processed signals served as input for various soundfield reconstruction algorithms, including the D-D approach [44], the non-parametric

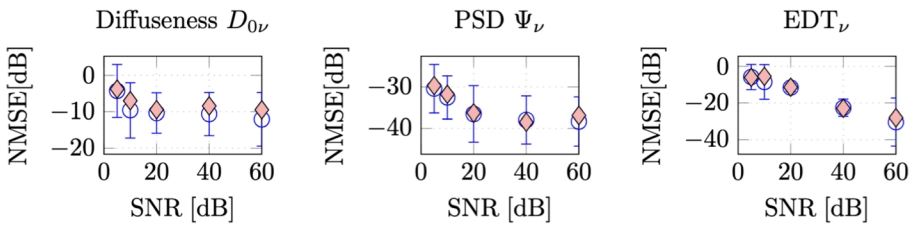

Fig. 6 Statistical distribution of the NMSE varying SNR levels. The blue circles represent the mean NMSE, with vertical error bars denoting ± one standard deviation. The red diamonds indicate the median values

reproducing kernel Hilbert space [38], that we denote as RKHS, and the proposed ParaDER algorithm. In total, we generated 200 signals, which were then mapped to the $A = 25$ virtual HOM locations, ensuring that a ground truth reference was available for each simulated HOM.

For playback, the reconstructed virtual HOM signals were filtered to derive second-order ambisonic tracks suitable for a virtual reality system. As described in [91], a total of 15 participants took part in the listening test, wearing a Meta Quest 2 visor and Sennheiser HD380 pro headphones amplified with Behringer HA400, in a 3D-rendered environment designed to accurately reproduce the real measured room from [49] as depicted in Fig. 7. For the subjective evaluation, the two sources were not active simultaneously. Participants were instructed to select and evaluate each source individually to provide isolated ratings for the perceptual attributes (Fig. 8).

Each participant rated five different models in a hidden manner:ParaDER (proposed method), D-D [44], RHKS [38], Ground Truth reference, 3.5kHz Low-pass filtered reference (mid. anchor) Following the methodology in [92], participants rated the stimuli on four perceptual attributes:

- **Overall audio quality**: the perceived fidelity and naturalness of the audio, considering artifacts, distortions, and clarity.
- **Localizability**: the accuracy with which a listener can determine the spatial position of sound sources within the auditory scene.
- **Spatial quality**: the sense of envelopment, depth, and width of the sound field, including spatial coherence and immersion.
- **Timbral quality**: the accuracy of tonal characteristics, including the preservation of frequency balance, harmonic integrity, and naturalness.

Moreover, during the evaluation the user can also switch the seat among the 25 different locations.

In Fig. 9, the results obtained from the voice simulations are presented. We observe that in all perceptual dimensions (audio quality, localizability, spatial quality,

Fig. 7 View of the simulated room according to [91]. Users can select each chair in the room, the sources are depicted as two yellow spheres

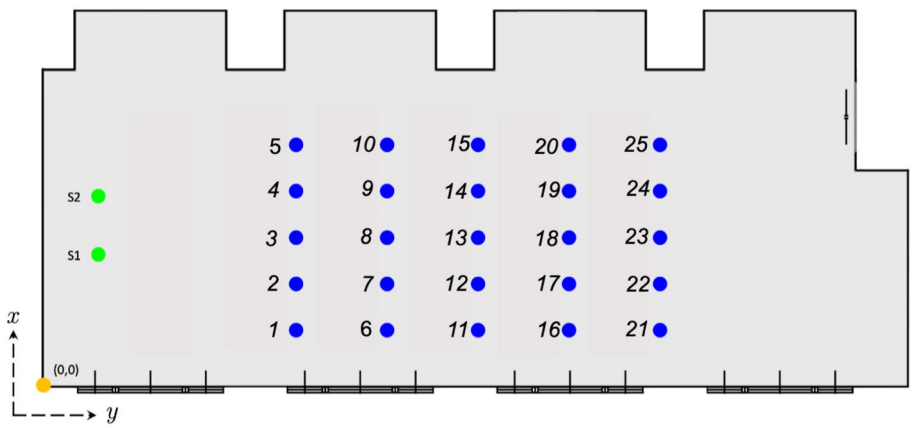

Fig. 8 Scheme of the room [49] used in [91] for the MUSHRA test

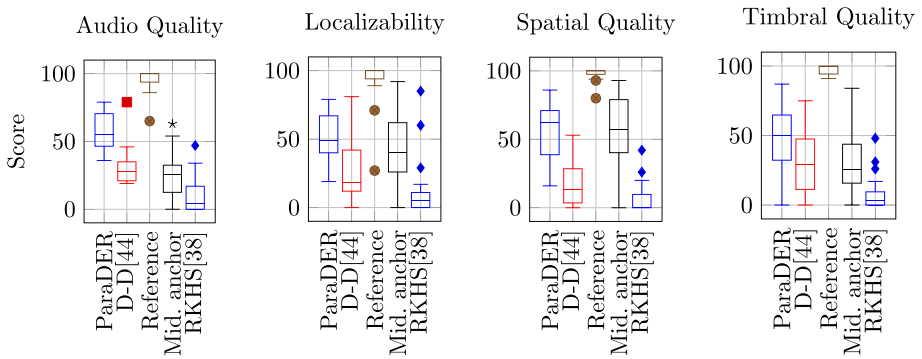

Fig. 9 Results of the evaluation of the quality metrics (audio quality, localizability, spatial quality, and timbral quality) of the MUSHRA test using female and male voice speech signal

and timbral quality) the proposed ParaDER algorithm consistently outperforms the competing methods.

While the perceptual scores are generally just above 50, this still reflects a clear advantage over the other methods and confirms the effectiveness of explicitly modeling ER. The moderate scores likely stem from the intrinsic complexity of human speech, which is rich in tonal detail and spatial cues. These results highlight both the robustness of the proposed approach and the remaining challenges in achieving perceptual transparency in realistic scenarios.

In contrast, Fig. 10 shows the results for music signal reconstruction. Here, the results of all reconstruction algorithms improve noticeably, with ParaDER achieving an average score of about 75, significantly higher than in the voice simulations. Indeed, studies on sound localization [93] have demonstrated that broader bandwidths and richer spectral content, characteristics typical of music, enhance localization performance. In contrast, speech signals, with their narrower frequency ranges and rapid temporal fluctuations, present more challenges for accurate spatial reconstruction.

Furthermore, focusing on specific perceptual aspects, such as localizability and spatial quality, the ParaDER method demonstrates a clear advantage over the D-D algorithm. This highlights the effectiveness of the ParaDER method in preserving the spatial characteristics of the sound field, enabling a more accurate and immersive experience for the listener. The significant improvement in localizability is particularly noteworthy, as it also confirms the importance of the diffuseness improvement shown in Fig. 5.

On the other hand, the RKHS-based algorithm performs poorly in the context of this study. Its scores remain consistently low across all perceptual classes, indicating that it is not well suited for the framework under consideration. Indeed, as discussed in Sect. 1, the algorithm's performance significantly deteriorates in high-frequency bands. This limitation is further exacerbated by the sparse microphone setup used in our scenario, which includes far fewer microphones than those employed in the original study [38]. As a result, the algorithm struggles to accurately approximate the sound field at the VM positions, cutting out the high frequency content and losing the spatiality as demonstrated by the MUSHRA test.

6 Conclusions and future works

In this work, we have presented a parametric virtual microphone framework that explicitly models direct, early, and diffuse components of a sound field, resulting

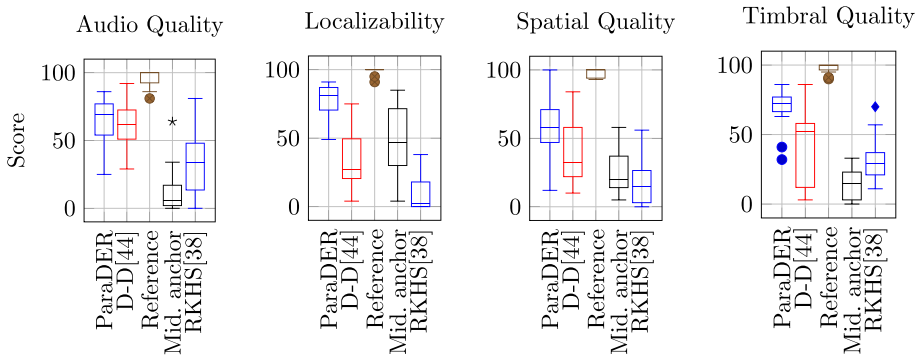

Fig. 10 Results of the evaluation of the quality metrics (audio quality, localizability, spatial quality, and timbral quality) of the MUSHRA test using music signal

in more accurate and perceptually realistic reconstructions in complex acoustic environments. The proposed method leverages a modular pipeline spanning virtual source estimation, time frequency filtering, and explicit early reflection modeling to achieve significant improvements in objective metrics such as NMSE, PSD reconstruction, diffuseness accuracy, and EDT. Our framework first localizes sources and estimates their signals via sparse regularized optimization of spherical harmonic coefficients. It then constructs virtual source parameters to propagate direct and early components to the virtual microphone, extracting the diffuse field as the residual. By fitting only zero-order (omnidirectional) coefficients for RIR retrieval, we maintain directivity patterns while simplifying the inversion. Finally, the synthesis module recombines these components, yielding high-fidelity spatial audio.

Results demonstrate that incorporating explicit ER modeling (ParaDER) yields over a 10dB improvement in NMSE for reflection reconstruction and consistently outperforms baseline D-D methods in both PSD synthesis and diffuseness accuracy across all harmonic orders. Listening tests further confirm that ParaDER delivers superior audio quality, spatial coherence, and timbral fidelity for both speech and music signals.

It is worth noting that in practical experiments the source positions are not perfectly known but estimated from the measured room impulse responses. We estimated that the average localization error in [49] data tests is about ≈ 25 cm, which is comparable to the physical size of the loudspeaker (≈ 40 cm). We did not observe a significant degradation in perceptual reconstruction quality, indicating that ParaDER is robust to realistic localization errors.

While the current implementation of ParaDER is not yet optimized for real-time operation, its parametric nature makes it inherently suitable for efficient, real-time applications. This potential is exemplified by its direct applicability to projects like the EU-funded REPERTORIUM, which aims to provide metaverse-ready classical music streaming and stands to benefit from ParaDER's low-microphone, real-time capabilities for immersive audio.

Future work will refine virtual-source estimation, potentially via adaptive PINNs, and extend the pipeline to dynamic sound fields, paving the way for live VR/AR experiences and large-scale cultural-heritage preservation.

Acknowledgements
We would like to express our sincere gratitude to Mr. Paolo Ostan and Ms. Francesca Del Gaudio for their valuable assistance in conducting the experimental test. This work has been funded by "REPERTORIUM project. Grant agreement number 101095065. Horizon Europe. Cluster II. Culture, Creativity and Inclusive Society. Call HORIZON-CL2-2022-HERITAGE-01-02" and by the European Union—Next Generation EU under the Italian National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.3, CUP D43C22003080001, partnership on "Telecommunications of the Future" (PE00000001—program "RESTART". We would also like to acknowledge the Multilayered Urban Sustainability Action (MUSA) project, which is funded by the European Union, for their contributions to this work.

Authors' contributions
Conceptualization, G.G. and M.P.; methodology, G.G., M.P. and M.P.; software, G.G. and M.P.; validation, G.G., M.P., F.A.; formal analysis, G.G. and M.P.; investigation, G.G.; data curation, G.G. and M.P.; writing—original draft preparation, G.G. and M.P.; writing—review and editing, G.G., M.P., F.A., A.S.; supervision, F.A. and A.S.; project administration, F.A. and A.S.; funding acquisition, F.A., A.S. All authors have read and agreed to the published version of the manuscript.

Data availability
This manuscript has no associated data.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 4 June 2025 Accepted: 3 November 2025
Published online: 04 December 2025

References

1. J. Yang, M. Wu, L. Han, A review of sound field control. Appl. Sci. **12**(14), (2022). https://doi.org/10.3390/app12147319
2. S.R. Quackenbush, J. Herre, Mpeg standards for compressed representation of immersive audio. Proc. IEEE **109**(9), 1578–1589 (2021). https://doi.org/10.1109/JPROC.2021.3075390
3. A.K. Bhowmik, Virtual and augmented reality: Human sensory-perceptual requirements and trends for immersive spatial computing experiences. J. Soc. Inf. Disp. **32**(8), 605–646 (2024). https://doi.org/10.1002/jsid.2001
4. J.G. Tylka, E.Y. Choueiri, in *Audio Engineering Society Convention*. Comparison of techniques for binaural navigation of higher-order ambisonic soundfields, (Audio Engineering Society, New York, NY, 2015) Paper 9421; Available from: https://aes2.org/publications/elibrary-page/?id=17977
5. L. Pisha, S. Yadegari, Specular path generation and near-reflective diffraction in interactive acoustical simulations. IEEE Trans. Vis. Comput. Graph. **30**(7), 3609–3621 (2024). https://doi.org/10.1109/TVCG.2023.3238662
6. H. Hacıhabiboğlu, E.D. Sena, Z. Cvetković, J.D. Johnston, J.O. Smith, Perceptual spatial audio recording, simulation, and rendering: An overview of spatial-audio techniques based on psychoacoustics. IEEE Signal Process. Mag. **34**, 36–54 (2017). https://doi.org/10.1109/MSP.2017.2666081
7. T. Sprunck, A. Deleforge, Y. Privat, C. Foy, Fully reversing the shoebox image source method: From impulse responses to room parameters. IEEE Trans. Audio Speech Lang. Process. **33**, 1023–1033 (2025). https://doi.org/10.1109/TASLPRO.2025.3536841
8. S. Damiano, F. Borra, A. Bernardini, F. Antonacci, A. Sarti, A compressive sensing approach for the reconstruction of the soundfield produced by directive sources in reverberant rooms. IEEE/ACM Trans. Audio Speech Lang. Process. **32**, 2667–2679 (2024). https://doi.org/10.1109/TASLP.2024.3398999
9. A. Politis, L. Pajunen, J. Leppänen, S. Mate, A. Eronen, in *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. Wide-area 6dof rendering of multi-point ambisonic recordings based on interpolation of spatial parameters (2023), pp. 1–5. https://doi.org/10.1109/WASPAA58266.2023.10248142
10. V. Pulkki, Spatial sound reproduction with directional audio coding. J. Audio Eng. Soc. **55**(6), 503–516 (2007)
11. J. Vilkamo, T. Lokki, V. Pulkki, Directional audio coding: Virtual microphone-based synthesis and subjective evaluation. J. Audio Eng. Soc. **57**(9), 709–724 (2009)
12. R. Schultz-Amling, F. Kuech, O. Thiergart, M. Kallinger, in *Audio Engineering Society Convention*. Acoustical zooming based on a parametric sound

field representation, vol. 128 (Audio Engineering Society, New York, NY, 2010)
13. S. Berge, N. Barrett, in *2nd International Symposium on Ambisonics and Spherical Acoustics*. High angular resolution planewave expansion (Proc. of the 2nd International Symposium on Ambisonics and Spherical Acoustics, Paris, France, 2010), pp. 6–7
14. G. Del Galdo, O. Thiergart, T. Weller, E.A.P. Habets, in *Hands-Free Speech Communication and Microphone Arrays (HSCMA)*. Generating virtual microphone signals using geometrical information gathered by distributed arrays (IEEE, Edinburgh, 2011), pp. 185–190
15. O. Thiergart, G. Del Galdo, M. Taseska, E.A.P. Habets, Geometry-based spatial sound acquisition using distributed microphone arrays. IEEE Trans. Audio Speech Lang. Process. **21**(12), 2583–2594 (2013)
16. K. Kowalczyk, O. Thiergart, M. Taseska, G. Del Galdo, V. Pulkki, E.A.P. Habets, Parametric spatial sound processing: A flexible and efficient solution to sound scene acquisition, modification, and reproduction. IEEE Signal Process. Mag. **32**(2), 31–42 (2015)
17. A. Plinge, S.J. Schlecht, O. Thiergart, T. Robotham, O. Rummukainen, E.A.P. Habets, in *Audio Engineering Society Conference: International Conference on Audio for Virtual and Augmented Reality*. Six-degrees-of-freedom binaural audio reproduction of first-order ambisonics with distance information (Audio Engineering Society, Redmond, WA, USA, 2018), p. 11
18. M. Pezzoli, F. Borra, F. Antonacci, A. Sarti, S. Tubaro, in *International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*. Estimation of the sound field at arbitrary positions in distributed microphone networks based on distributed ray space transform (IEEE, Piscataway, NJ, 2018), pp. 186–190.
19. M. Pezzoli, F. Borra, F. Antonacci, A. Sarti, S. Tubaro, in *26th European Signal Processing Conference (EUSIPCO)*. Reconstruction of the virtual microphone signal based on the distributed ray space transform (IEEE, Piscataway, NJ, 2018), pp. 1537–1541.
20. P. Samarasinghe, T.D. Abhayapala, M.A. Poletti, in *International Workshop on Acoustic Signal Enhancement (IWAENC)*. 3D spatial soundfield recording over large regions (Institute of Electrical and Electronics Engineers Inc., Aachen, Germany, 2012), pp. 1–4
21. P. Samarasinghe, T.D. Abhayapala, M.A. Poletti, Wavefield analysis over large areas using distributed higher order microphones. IEEE/ACM Trans. Audio Speech Lang. Process. **22**(3), 647–658 (2014)
22. J.G. Tylka, E.Y. Choueiri, in *Audio Engineering Society Conference: International Conference on Audio for Virtual and Augmented Reality*. Soundfield navigation using an array of higher-order ambisonics microphones (Audio Engineering Society, 2016), p. 10. https://doi.org/10.17743/aesconf.2016.978-1-942220-10-7
23. N. Ueno, S. Koyama, H. Saruwatari, Sound field recording using distributed microphones based on harmonic analysis of infinite order. IEEE Signal Process. Lett. **25**(1), 135–139 (2018). https://doi.org/10.1109/LSP.2017.2774351
24. Y. Takida, S. Koyama, H. Saruwataril, in *26th European Signal Processing Conference, (EUSIPCO)*. Exterior and interior sound field separation using convex optimization: Comparison of signal models (IEEE , 2018), pp. 2549–2553. https://doi.org/10.23919/EUSIPCO.2018.8553535
25. S. Koyama, L. Daudet, Sparse representation of a spatial sound field in a reverberant environment. IEEE J. Sel. Top. Signal Process. **13**(1), 172–184 (2019)
26. F. Borra, I.D. Gebru, D. Marković, in *International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*. Soundfield reconstruction in reverberant environments using higher-order microphones and impulse response measurements (IEEE, Piscataway, NJ, 2019), pp. 281–285.
27. F. Borra, I.D. Gebru, D. Marković, in *Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. 1st-order microphone array system for large area sound field recording and reconstruction: Discussion and preliminary results (IEEE, 2019), p. 5
28. S. Damiano, F. Miotello, M. Pezzoli, A. Bernardini, F. Antonacci, A. Sarti, T. Van Waterschoot, in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. A zero-shot physics-informed dictionary learning approach for sound field reconstruction (IEEE, Piscataway, NJ, 2025), pp. 1–5. https://doi.org/10.1109/ICASSP49660.2025.10888108
29. Y.J. Wu, T.D. Abhayapala, Spatial multizone soundfield reproduction: Theory and design. Trans. Audio Speech Lang. Proc. **19**(6), 1711–1720 (2011). https://doi.org/10.1109/TASL.2010.2097249
30. T. Betlehem, T.D. Abhayapala, Theory and design of sound field reproduction in reverberant rooms. J. Acoust. Soc. Am. **117**(4), 2100–2111 (2005). https://doi.org/10.1121/1.1863032
31. W. Zhang, P.N. Samarasinghe, H. Chen, T.D. Abhayapala, Surround by sound: A review of spatial audio recording and reproduction. Appl. Sci. **7**(5) (2017). https://doi.org/10.3390/app7050532
32. D.N. Zotkin, R. Duraiswami, N.A. Gumerov, in *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. Regularized hrtf fitting using spherical harmonics (2009), pp. 257–260. https://doi.org/10.1109/ASPAA.2009.5346521
33. H. Liu, Y. Fang, Q. Huang, Efficient representation of head-related transfer functions with combination of spherical harmonics and spherical wavelets. IEEE Access **7**, 78214–78222 (2019). https://doi.org/10.1109/ACCESS.2019.2921388
34. F. Lluís, P. Martínez-Nuevo, M. Bo Møller, S. Ewan Shepstone, Sound field reconstruction in rooms: Inpainting meets super-resolution. J. Acoust. Soc. Am. **148**(2), 649–659 (2020). https://doi.org/10.1121/10.0001687
35. S. Damiano, F. Borra, A. Bernardini, F. Antonacci, A. Sarti, in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. Soundfield reconstruction in reverberant rooms based on compressive sensing and image-source models of early reflections (2021), pp. 366–370. https://doi.org/10.1109/WASPAA52581.2021.9632746
36. S. Koyama, H. Saruwatari, in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Sound field decomposition in reverberant environment using sparse and low-rank signal models (IEEE, Piscataway, NJ, 2016)
37. M. Pezzoli, M. Cobos, F. Antonacci, A. Sarti, in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Sparsity-based sound field separation in the spherical harmonics domain (2022), pp. 1051–1055. https://doi.org/10.1109/ICASSP43922.2022.9746391
38. J.G.C. Ribeiro, S. Koyama, R. Horiuchi, H. Saruwatari, Sound field estimation based on physics-constrained kernel interpolation adapted to environment. IEEE/ACM Trans. Audio Speech Lang. Process. **32**, 4369–4383 (2024). https://doi.org/10.1109/TASLP.2024.3467951
39. A. Figueroa-Duran, E. Fernandez-Grande, Reconstruction of reverberant sound fields over large spatial domains. J. Acoust. Soc. Am. **157**(1), 180–190 (2025). https://doi.org/10.1121/10.0034833
40. F. Ronchini, L. Comanducci, M. Pezzoli, F. Antonacci, A. Sarti, in *2024 32nd European Signal Processing Conference (EUSIPCO)*. Room transfer function reconstruction using complex-valued neural networks and irregularly distributed microphones (2024), pp. 441–445. https://doi.org/10.23919/EUSIPCO63174.2024.10715145
41. F. Miotello, L. Comanducci, M. Pezzoli, A. Bernardini, F. Antonacci, A. Sarti, in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Reconstruction of sound field through diffusion models (2024), pp. 1476–1480. https://doi.org/10.1109/ICASSP48485.2024.10446761
42. M. Raissi, P. Perdikaris, G.E. Karniadakis, Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. J. Comput. Phys. **378**, 686–707 (2019). https://doi.org/10.1016/j.jcp.2018.10.045
43. S. Koyama, J.G.C. Ribeiro, T. Nakamura, N. Ueno, M. Pezzoli, Physics-informed machine learning for sound field estimation: Fundamentals, state of the art, and challenges. IEEE Signal Process. Mag. **41**(6), 60–71 (2024). https://doi.org/10.1109/MSP.2024.3465896
44. M. Pezzoli, F. Borra, F. Antonacci, S. Tubaro, A. Sarti, A parametric approach to virtual miking for sources of arbitrary directivity. IEEE/ACM Trans. Audio Speech Lang. Process. **28**, 2333–2348 (2020). https://doi.org/10.1109/TASLP.2020.3012058
45. D. Albertini, G. Greco, A. Bernardini, A. Sarti, in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Diffusion-based sound source localization using networks of planar microphone arrays (2023), pp. 1–5. https://doi.org/10.1109/ICASSP49357.2023.10095405
46. D. Albertini, A. Bernardini, G. Greco, A. Sarti, Diffusion-based sound source localization using a distributed network of microphone arrays. Sensors **25**(7) (2025). https://doi.org/10.3390/s25072078
47. A.M. Molaei, B. Zakeri, S.M.H. Andargoli, M.A.B. Abbasi, V. Fusco, O. Yurduseven, A comprehensive review of direction-of-arrival estimation and localization approaches in mixed-field sources scenario. IEEE Access **12**, 65883–65918 (2024). https://doi.org/10.1109/ACCESS.2024.3398351
48. Y. Sumura, D.D. Carlo, A. Arie Nugraha, Y. Bando, K. Yoshii, in *2024 18th International Workshop on Acoustic Signal Enhancement (IWAENC)*. Joint

audio source localization and separation with distributed microphone arrays based on spatially-regularized multichannel nmf (2024), pp. 145–149. https://doi.org/10.1109/IWAENC61483.2024.10694042
49. F. Miotello, P. Ostan, M. Pezzoli, L. Comanducci, A. Bernardini, F. Antonacci, A. Sarti, in *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*. Homula-rir: A room impulse response dataset for teleconferencing and spatial audio applications acquired through higher-order microphones and uniform linear microphone arrays (2024), pp. 795–799. https://doi.org/10.1109/ICASSPW62465.2024.10626753
50. A. Schwarz, W. Kellermann, Coherent-to-diffuse power ratio estimation for dereverberation. IEEE Trans. Audio Speech Lang. Process. **23**(6), 1006–1018 (2015)
51. J.B. Allen, D.A. Berkley, Image method for efficiently simulating small-room acoustics. J. Acoust. Soc. Am. **65**(4), 943–950 (1979)
52. B. Loesch, B. Yang, in *International Workshop on Acoustic Signal Enhancement (IWAENC)*. Source number estimation and clustering for underdetermined blind source separatio (IEEE, Piscataway, NJ, 2008).
53. S. Araki, T. Nakatani, H. Sawada, S. Makino, in *Independent Component Analysis and Signal Separation*, ed by T. Adali, C. Jutten, J.M.T. Romano, A.K. Barros. Stereo source separation and source counting with map estimation with Dirichlet prior considering spatial aliasing problem (Springer, Berlin, 2009), pp. 742–750
54. S. Arberet, R. Gribonval, F. Bimbot, A robust method to count and locate audio sources in a multichannel underdetermined mixture. IEEE Trans. Signal Process. **58**(1), 121–133 (2010)
55. D. Pavlidi, A. Griffin, M. Puigt, A. Mouchtaris, in *Sensor Array and Multichannel Signal Processing Workshop (SAM)*. Source counting in real-time sound source localization using a circular microphone array (IEEE, 2012), pp. 521–524. IEEE 445 Hoes Lane Piscataway, NJ 08854-4141 USA
56. O. Walter, L. Drude, R. Haeb-Umbach, in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Source counting in speech mixtures by nonparametric bayesian estimation of an infinite gaussian mixture model (IEEE, Piscataway, NJ, 2015), pp. 459–463
57. S. Pasha, J. Donley, C. Ritz, in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. Blind speaker counting in highly reverberant environments by clustering coherence features (2017), pp. 1684–1687
58. C. Xu, S. Li, G. Liu, Y. Zhang, E. Miluzzo, Y.-F. Chen, J. Li, B. Firner, in *ACM International Joint Conference on Pervasive and Ubiquitous Computing*. Crowd++: Unsupervised speaker count with smartphones, UbiComp'13 (Association for Computing Machinery, New York, 2013), pp. 43–52. https://doi.org/10.1145/2493432.2493435
59. F. Stöter, S. Chakrabarty, B. Edler, E.A.P. Habets, Countnet: Estimating the number of concurrent speakers using supervised learning. IEEE/ACM Trans. Audio Speech Lang. Process. **27**(2), 268–282 (2019)
60. M. Cobos, F. Antonacci, A. Alexandridis, A. Mouchtaris, B. Lee, A survey of sound source localization methods in wireless acoustic sensor networks. Wirel. Commun. Mob. Comput. **2017** (Wiley,2017) pp.1–24 https://doi.org/10.1155/2017/3956282
61. X. Sheng, Y.-H. Hu, Maximum likelihood multiple-source localization using acoustic energy measurements with wireless sensor networks. IEEE Trans. Signal Process. **53**(1), 44–53 (2005)
62. A. Canclini, F. Antonacci, A. Sarti, S. Tubaro, Acoustic source localization with distributed asynchronous microphone networks. IEEE Trans. Audio Speech Lang. Process. **21**(2), 439–443 (2013)
63. F. Borra, M. Pezzoli, L. Comanducci, A. Bernardini, F. Antonacci, S. Tubaro, A. Sarti, in *28th European Signal Processing Conference (EUSIPCO)*. A fast ray space transform for wave field processing using acoustic arrays (IEEE, Piscataway, NJ, 2020)
64. DiBiase, J.H.: A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays. PhD thesis (Brown University, 2000)
65. H. Teutsch, W. Kellermann, in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. Acoustic source detection and localization based on wavefield decomposition using circular microphone arrays (IEEE, Piscataway, NJ, 2008), pp. 273–276.
66. S. Adavanne, A. Politis, J. Nikunen, T. Virtanen, Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. IEEE J. Sel. Top. Signal Process. **13**(1), 34–48 (2019)
67. L. Chen, G. Chen, L. Huang, Y.-S. Choy, W. Sun, Multiple sound source localization, separation, and reconstruction by microphone array: A dnn-based approach. Appl. Sci. **12**(7), 3428 (2022)
68. Y. Gong, S. Liu, X.-L. Zhang, *APSIPA Annual Summit and Conference.* End-to-end two-dimensional sound source localization with ad-hoc microphone arrays. (Proceedings of 2022 APSIPA Annual Summit and Conference, Chiang Mai, Thailand, 2022).
69. G. Greco, S. Messana, M. Pezzoli, M. Cobos, F. Antonacci, in *2025 34nd European Signal Processing Conference (EUSIPCO)*. Dereverberation of relative harmonic coefficients via CNNs for acoustic source DOA estimation (IEEE, Palermo, 2025)
70. M. Miyoshi, Y. Kaneda, Inverse filtering of room acoustics. IEEE Trans. Acoust. Speech Signal Process. **36**(2), 145–152 (1988)
71. M. Delcroix, T. Hikichi, M. Miyoshi, Precise dereverberation using multichannel linear prediction. IEEE Trans. Audio Speech Lang. Process. **15**(2), 430–440 (2007)
72. H. Buchner, R. Aichner, W. Kellermann, in *International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*. TRINICON: A versatile framework for multichannel blind signal processing, vol. 3 (IEEE, Piscataway, NJ, 2004), p. 889.
73. O. Thiergart, G. Del Galdo, E.A.P. Habets, On the spatial coherence in mixed sound fields and its application to signal-to-diffuse ratio estimation. J. Acoust. Soc. Am. **132**(4), 2337–2346 (2012)
74. S.V. Vaseghi, *Advanced Digital Signal Processing and Noise Reduction* (Wiley, Hoboken, NJ, 2008)
75. J. Sohn, N.S. Kim, W. Sung, A statistical model-based voice activity detection. IEEE Signal Process. Lett. **6**(1), 1–3 (1999)
76. R. Zelinski, in *International Conference on Acoustics, Speech, and Signal Processing, (ICASSP)*. A microphone array with adaptive post-filtering for noise reduction in reverberant rooms, vol. 5, (IEEE, Piscataway, NJ, 1988), pp. 2578–2581.
77. I.A. McCowan, H. Bourlard, Microphone array post-filter based on noise field coherence. IEEE Trans. Speech Audio Process. **11**(6), 709–716 (2003)
78. R.J. Polge, E.M. Mitchell, Impulse response determination by cross correlation. IEEE Trans. Aerosp. Electron. Syst. **AES–6**(1), 91–97 (1970). https://doi.org/10.1109/TAES.1970.310015
79. J. Cao, Z. Yang, X. Chen, R. Yan, From pseudo to real: Generalized subspace method for power spectrum reconstruction. IEEE Trans. Ind. Electron. **71**(4), 4141–4150 (2024). https://doi.org/10.1109/TIE.2023.3279569
80. M. Ester, H.-P. Kriegel, J. Sander, X. Xu, in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD)*. A density-based algorithm for discovering clusters in large spatial databases with noise (AAAI Press, Portland, 1996), pp. 226–231
81. L. Madmoni, Z. Ben-Hur, J. Donley, V. Tourbabin, B. Rafaely, Design and analysis of binaural signal matching with arbitrary microphone arrays and listener head rotations. EURASIP J. Audio Speech Music Process. **2025**(11), 1–18 (2025)
82. N. Gößling, D. Marquardt, S. Doclo, Perceptual evaluation of binaural mvdr-based algorithms to preserve the interaural coherence of diffuse noise fields. Trends Hear. **24**, 2331216520919573 (2020)
83. Hardin, R.H., Sloane, N.J.A. McLaren's improved snub cube and other new spherical designs in three dimensions. Discrete Comput Geom 15, 429–441 (SpringerNature, 1996). https://doi.org/10.1007/BF02711518
84. European Broadcasting Union: Sound quality assessment material recording for subjective tests. Technical report, European Broadcasting Union (2008). European Broadcasting Union.(Accessed: February 10, 2025) https://tech.ebu.ch/publications/sqamcd
85. Habets, E.A.P.: Room impulse response generator. Technical Report 2.4, Technische Universiteit Eindhoven, Tech. Rep (2006)
86. N. Epain, C.T. Jin, Spherical harmonic signal covariance and sound field diffuseness. IEEE/ACM Trans. Audio Speech Lang. Proc. **24**(10), 1796–1807 (2016). https://doi.org/10.1109/TASLP.2016.2585862
87. J.S. Bradley, Review of objective room acoustics measures and future needs. Appl. Acoust. **72**(10), 713–720 (2011). https://doi.org/10.1016/j.apacoust.2011.04.004
88. T.D. Abhayapala, D.B. Ward, in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Theory and design of high order sound field microphones using spherical microphone array, vol. 2 (2002), pp. 1949–1952. https://doi.org/10.1109/ICASSP.2002.5745011

89. J. Richter, Y.-C. Wu, S. Krenn, S. Welker, B. Lay, S. Watanabe, A. Richard, T. Gerkmann, in *ISCA Interspeech*. EARS: An anechoic fullband speech dataset benchmarked for speech enhancement and dereverberation (Proc. Interspeech, Kos, Grece, 2024), pp. 4873–4877. https://doi.org/10.21437/Interspeech.2024-1532024
90. B. Li, X. Liu, K. Dinesh, Z. Duan, G. Sharma, Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications. IEEE Trans. Multimed. **21**(2), 522–535 (2019). https://doi.org/10.1109/TMM.2018.2856090
91. P. Ostan, F.D. Gaudio, F. Miotello, M. Pezzoli, F. Antonacci, in *Proceedings of the Forum Acusticum*. VR-PTOLEMAIC: A virtual environment for the perceptual testing of spatial audio algorithms, vol 2025 (EAA, Malaga, Spain, 2025)
92. O.S. Rummukainen, T. Robotham, S.J. Schlecht, A. Plinge, J. Herre, E.A.P. Habets, in *AES International Conference on Audio for Virtual and Augmented Reality*. Audio quality evaluation in virtual reality: Multiple stimulus ranking with behavior tracking (2018), pp. 1–10. (Accessed: February 10, 2025) https://www.aes.org/e-lib/browse.cfm?elib=19678
93. A. Carlini, C. Bordeau, M. Ambard, Auditory localization: a comprehensive practical review. Front. Psychol. **Volume 15 - 2024** (2024). https://doi.org/10.3389/fpsyg.2024.1408073

Publisher's Note