Signal Processing 249 (2026) 110774

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Iterative high-order spherical harmonic diagonal-unloading beamforming for localizing direct sound and dominant reflections in reverberant rooms

Daniele Salvati

Department of Mathematics, Computer Science and Physics, University of Udine, via delle Scienze, 206, Udine, 33100, Italy

ARTICLE INFO

Keywords:
Spherical microphone arrays
Spherical harmonics
Iterative diagonal-unloading
High-resolution beamforming
Early reflections
Reverberation
Direction-of-arrival estimation
Room acoustics

ABSTRACT

Reverberation and early reflections create spurious peaks in acoustic power maps and bias direction-of-arrival (DOA) estimates. At the same time, the direct path and dominant reflections carry useful information about the acoustic environment. This paper investigates joint localization of direct sound and multiple dominant reflected components in single-source reverberant environments using spherical microphone arrays (SMAs). It introduces an iterative spherical-harmonic (SH) diagonal-unloading (DU) beamforming framework for multipath analysis. Microphone signals are encoded in the SH domain and fused into a broadband covariance matrix by frequency smoothing with per-bin trace normalization. Multi-peak localization alternates DU map evaluation and iterative covariance filtering, with detected directions suppressed through a two-sided null-steering update and per-iteration trace renormalization. Performance is assessed through controlled simulations spanning reverberation time, signal-to-noise ratio (SNR), signal-to-diffuse-noise ratio (SDNR), and gain mismatch, together with experiments based on multichannel room impulse responses recorded with a spherical microphone array. Results in terms of root-mean-square error (RMSE) and accuracy rate (AR) show that increasing SH order improves localization of the direct path and strongest reflected components, whereas later iterations become more challenging as weaker components are extracted. The proposed method outperforms the baselines while retaining a computationally attractive structure for SH processing.

1. Introduction

Accurate localization of acoustic sources in enclosed environments remains challenging because of multipath propagation and reverberation. In typical rooms, boundary reflections create multiple propagation paths whose energy can be comparable to that of the direct sound and whose arrivals may fall within the same analysis window. As a result, acoustic maps often exhibit multiple peaks that do not uniquely identify the direct path, which biases direction-of-arrival (DOA) estimates and reduces localization accuracy. This issue is particularly important for high-angular-resolution methods, which can resolve multiple components but may also emphasize reflections when they are coherent with the direct sound.

These challenges have motivated extensive research on robust localization in reverberant conditions, where reflections are usually treated as interference to be suppressed or mitigated. Representative examples include robust beamforming, spatial filtering, and neural methods developed in recent years [1–11].

Early reflections, however, have increasingly been recognized as informative cues about the acoustic environment. Joint estimation of the direct path and the directions of dominant reflected components can reveal properties of reflective boundaries, room geometry, and propagation structure, thereby supporting room-acoustics analysis and geometry-related inference. From this viewpoint, reflections are not merely interference to be suppressed but are informative components of the acoustic scene that can be explicitly analyzed and exploited [12].

A large portion of the reflection-aware literature has been developed within a room impulse response (RIR) analysis framework. In this setting, measured or estimated RIRs are analyzed to extract reflection time-of-arrival (TOA) and DOA information. Reflection TOAs can be converted into geometric constraints for planar-reflector localization, for example by using Hough-transform-based ambiguity resolution to improve robustness in noise [13]. Time–frequency directional analysis of measured RIRs with spherical microphone arrays (SMAs) has been combined with spherical-harmonic (SH) processing, delay-and-sum beamforming, and wavelet-based peak detection to estimate the time and direction of early reflections [14]. Subsequent studies addressed room-geometry inference by localizing reflecting boundaries from reflection TOA/DOA information, including approaches that handle both convex and non-convex room shapes [15]. More recently, learning-based strategies have been proposed to regress room dimensions and reflection coefficients directly from RIRs, including settings that do not require known

E-mail address: daniele.salvati@uniud.it

https://doi.org/10.1016/j.sigpro.2026.110774
Received 18 February 2026; Received in revised form 16 April 2026; Accepted 15 June 2026
Available online 16 June 2026

source–receiver positions and operate from a single RIR measurement [16].

Complementary to explicit RIR analysis, another line of research estimates direct and reflected components more directly from microphone signals, or from short RIR windows, using array-processing formulations tailored to coherent multipath. Maximum-likelihood approaches have been shown to be robust to noise when applied to coherent reflections measured with SMAs [17]. In a related direction, room-geometry inference has been pursued by jointly estimating the DOAs of the direct path and early reflections using high-resolution broadband beamforming on SMAs and combining these cues through geometric constraints [18]. More recently, blind early-reflection DOA estimation without the source signal or explicit RIR knowledge has been advanced through phase-aligned correlation methods [19]. In parallel, recent 3-D sound event localization and detection studies have shown that combining spatial-coherence cues with direct-path-dominance features improves the robustness of distance estimation in real-world reverberant scenes, further highlighting the value of separating direct-path and reflected energy for spatial inference [20].

Despite these advances, many methods still focus on the direct path and only a small number of early reflections. Reliable separation becomes increasingly difficult as the number of components grows under coherent single-source multipath, limited array aperture, and progressively weaker reflected energy. As a result, scalable formulations that support deep multi-peak extraction while remaining computationally tractable are still relatively limited.

This paper addresses estimation of the DOAs of the direct path and of a relatively large set of dominant propagation components generated by a single source and observed by an SMA. Although the extracted components may extend beyond strictly defined early reflections, they represent the dominant coherent propagation paths within the analysis window. The proposed approach operates in the SH domain and employs diagonal-unloading (DU) beamforming [21,22]. Relative to eigenspace-based high-resolution methods, such as multiple signal classification (MUSIC) [23], and inversion-based formulations related to minimum variance distortionless response (MVDR) [24], DU is computationally attractive because map evaluation requires neither explicit eigendecomposition nor matrix inversion. A single broadband SH covariance matrix is formed through frequency smoothing with per-bin trace normalization [25]. Multi-peak extraction then alternates DU map evaluation with iterative covariance filtering. At each iteration, the dominant steering direction is detected, a two-sided null-steering operator suppresses the corresponding component, and the updated covariance is trace-normalized before the next iteration. This strategy is related in spirit to deflation-based and CLEAN-like approaches for coherent component separation [26]. A similar iterative suppression principle has been explored in compact-array DU processing [27]. In the present work, this principle is developed within a broadband SH-domain framework for localizing direct sound and dominant reflections with SMAs.

The main contributions are as follows:

- An SH-domain iterative DU framework for sequential localization of the direct path and a comparatively large set of dominant component DOAs under coherent single-source multipath;
- A broadband formulation based on a single SH covariance matrix together with a two-sided null-steering covariance update and per-iteration trace renormalization for repeated multi-peak extraction;
- A systematic evaluation across SH order, extraction depth, reverberation, signal-to-noise ratio (SNR), signal-to-diffuse-noise ratio (SDNR), gain mismatch, and measured SMA data, using root-mean-square error (RMSE) and accuracy rate (AR) to assess performance.

The approach is evaluated through controlled reverberant simulations and measured-data experiments on the public Arni spatial room dataset [28], which provides real multichannel room impulse responses recorded with a 32-channel SMA in a reverberant room. The experimental study compares the proposed method with SH-DU [25], SH-MUSIC [23,29], a standard SH-domain steered-response power (SRP) implementation [30], and microphone-domain SRP with phase transform (PHAT) [31]. The results show that increasing the SH order improves localization of the direct path and the strongest reflected components, whereas later iterations become progressively more difficult as weaker and more ambiguous components are extracted. Overall, the study highlights the potential of iterative SH-domain DU processing for structured localization of dominant acoustic components in reverberant rooms.

2. Signal model

An SMA composed of Q omnidirectional microphones is considered, with sensor positions $\mathbf{r}_q = [x_q, y_q, z_q]^\top \in \mathbb{R}^3$, $q = 1, \dots, Q$, and array center $\mathbf{r}_c = [x_c, y_c, z_c]^\top$. A single acoustic source is assumed to be active in a reverberant enclosure. The sound field is modeled as the direct component plus a finite set of dominant reflected components generated by the surrounding boundaries. This representation emphasizes the coherent propagation paths that are most relevant to the directional analysis performed by the array.

Each propagation path is characterized by a DOA

$$\boldsymbol{\Omega}_p = [\theta_p,\ \vartheta_p]^\top, \tag{1}$$

where $\theta_p \in [0, 2\pi)$ denotes azimuth and $\vartheta_p \in [0, \pi]$ denotes colatitude (polar angle).

2.1. STFT-domain array model

Let $(\cdot)(k, f)$ denote the short-time Fourier transform (STFT) coefficient at time-frame index k and frequency-bin index f. The center frequency associated with bin f is denoted by ν_f. In the adopted formulation, each dominant propagation component is represented at the array by an equivalent plane-wave contribution. The array response therefore depends only on its direction of arrival. This approximation is used for both the direct path and the dominant reflected components considered in the analysis. The STFT-domain signal at the qth microphone is

$$x_q(k, f) = \sum_{p=0}^{P} a_q(f, \boldsymbol{\Omega}_p)\, s_p(k, f) + v_q(k, f), \tag{2}$$

where $p = 0$ denotes the direct component and $p \geq 1$ denotes the reflected components. The term $s_p(k, f)$ denotes the STFT coefficient associated with the pth component, and $v_q(k, f)$ is an additive residual term that collects sensor noise and contributions not explicitly represented by the dominant propagation components. The scalar $a_q(f, \boldsymbol{\Omega}_p)$ denotes the array response of the qth microphone to a plane wave impinging from $\boldsymbol{\Omega}_p$.

Define the microphone-signal vector as

$$\mathbf{x}(k, f) = \begin{bmatrix} x_1(k, f) & \cdots & x_Q(k, f) \end{bmatrix}^\top. \tag{3}$$

Accordingly, (2) can be written as

$$\mathbf{x}(k, f) = \sum_{p=0}^{P} \mathbf{a}(f, \boldsymbol{\Omega}_p)\, s_p(k, f) + \mathbf{v}(k, f), \tag{4}$$

where $\mathbf{v}(k, f) = [v_1(k, f), \dots, v_Q(k, f)]^\top$ and

$$\mathbf{a}(f, \boldsymbol{\Omega}_p) = [a_1(f, \boldsymbol{\Omega}_p), \dots, a_Q(f, \boldsymbol{\Omega}_p)]^\top \tag{5}$$

is the steering vector associated with direction $\boldsymbol{\Omega}_p$.

2.2. Spherical harmonic domain representation

To exploit the spatial structure of SMAs, the microphone signals are encoded into the SH domain up to order L [32]. The number of retained SH components is $N_{\mathrm{SH}} = (L + 1)^2$, and it is assumed that

$$(L + 1)^2 \leq Q, \tag{6}$$

so that the corresponding SH coefficients can be estimated from the microphone signals. In the following, L denotes the SH order retained for localization, whereas $\Lambda \geq L$ denotes the truncation order used in the transfer model for regularized encoding.

Let $\mathbf{E}_{L,\Lambda}(f) \in \mathbb{C}^{(L+1)^2 \times Q}$ denote the SH encoding matrix. It returns the order-L coefficients while relying on a transfer model truncated at order Λ. The SH-domain signal is defined as

$$\mathbf{b}(k, f) = \mathbf{E}_{L,\Lambda}(f)\,\mathbf{x}(k, f). \tag{7}$$

In practice, $\mathbf{E}_{L,\Lambda}(f)$ is obtained by regularized inversion of the SMA transfer matrix. This improves numerical stability and limits noise amplification at ill-conditioned frequencies [33].

Let $\boldsymbol{\Omega}_q = [\theta_q, \vartheta_q]^\top$ denote the angular position of the qth microphone on a sphere of radius R. By truncating the SH expansion to order Λ, the microphone-domain observations satisfy

$$\mathbf{x}(k, f) \approx \mathbf{H}_\Lambda(f)\,\mathbf{b}_\Lambda(k, f), \tag{8}$$

where $\mathbf{b}_\Lambda(k, f) \in \mathbb{C}^{(\Lambda+1)^2}$ is the vector collecting the spherical-harmonic coefficients of the sound field, truncated to order Λ, and $\mathbf{H}_\Lambda(f) \in \mathbb{C}^{Q\times(\Lambda+1)^2}$ has entries

$$\big[\mathbf{H}_\Lambda(f)\big]_{q,(l,m)} = w_l\left(\frac{2\pi v_f R}{c}\right) Y_l^m\big(\theta_q, \vartheta_q\big), \qquad 0 \leq l \leq \Lambda,\ -l \leq m \leq l, \tag{9}$$

with $w_l(\cdot)$ the modal strength and $Y_l^m(\cdot)$ the spherical harmonic function. The order-L transfer matrix $\mathbf{H}_L(f) \in \mathbb{C}^{Q\times(L+1)^2}$ is obtained by retaining from $\mathbf{H}_\Lambda(f)$ the columns corresponding to $0 \leq l \leq L$. Here, c denotes the speed of sound.

In this implementation, the encoding matrix is computed via a Tikhonov-regularized left pseudo-inverse [34],

$$\mathbf{E}_{L,\Lambda}(f) = \mathbf{H}_L^{\mathrm{H}}(f)\Big(\mathbf{H}_\Lambda(f)\mathbf{H}_\Lambda^{\mathrm{H}}(f) + [\beta(f)]^2\mathbf{I}_Q\Big)^{-1}, \tag{10}$$

where $\mathbf{I}_Q$ is the $Q \times Q$ identity matrix. The regularization is selected adaptively as

$$\beta(f) = \beta_{\mathrm{set}}\ \sigma_{\max}\big(\mathbf{H}_\Lambda(f)\big), \tag{11}$$

with $\sigma_{\max}(\cdot)$ the largest singular value and $\beta_{\mathrm{set}} > 0$ a dimensionless scaling constant [35]. The factor β_{set} controls the overall amount of regularization in the SH encoding stage. Through (11), the regularization remains frequency-adaptive because it scales with the largest singular value of the truncated transfer matrix at each frequency. At the same time, β_{set} acts as a global dimensionless stabilization factor that balances inversion accuracy, numerical robustness, and mitigation of noise amplification at ill-conditioned frequencies. In principle, the truncation order may be frequency-dependent. In the present implementation, however, a single modeling order Λ is fixed from the highest analyzed frequency, namely

$$\Lambda = \left\lceil \frac{2\pi v_{f_{\max}} R}{c} \right\rceil, \tag{12}$$

so that the same regularized encoding structure is used across the analyzed frequency band.

Under the standard SH-domain plane-wave model, (4) becomes

$$\mathbf{b}(k, f) = \sum_{p=0}^{P} \mathbf{y}(\boldsymbol{\Omega}_p)\, s_p(k, f) + \mathbf{n}(k, f), \tag{13}$$

where $\mathbf{y}(\boldsymbol{\Omega}_p) \in \mathbb{C}^{(L+1)^2}$ denotes the SH steering vector associated with direction $\boldsymbol{\Omega}_p$, and $\mathbf{n}(k, f)$ denotes the encoded residual term produced by the regularized SH encoding of $\mathbf{v}(k, f)$. A key property used in this work is that $\mathbf{y}(\boldsymbol{\Omega}_p)$ is frequency-independent. This property motivates the use of broadband SH-domain spatial statistics without introducing frequency-dependent steering vectors.

2.3. Problem formulation and challenges

The objective is to estimate the DOAs of the direct component and of a set of dominant propagation components,

$$\mathcal{S} = \{\boldsymbol{\Omega}_0, \boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_P\}. \tag{14}$$

Reflected components are indexed by their dominance in the observed sound field, i.e., by the extraction order of the algorithm, rather than by geometric reflection order. This distinction is important in reverberant environments, where several reflections may have comparable energy and may arrive from closely spaced directions.

Several factors make this problem challenging. Direct sound and reflections are coherent because they originate from the same emitted signal, which limits the effectiveness of methods based on source decorrelation. The angular resolution of an SMA is governed by the retained SH order L, and therefore by the number of microphones and the effective aperture. Low-order processing yields broad spatial responses that merge nearby components and can propagate errors to later extractions. As extraction depth increases, progressively weaker components are sought after repeated suppression steps, making covariance scaling and numerical conditioning increasingly important. In addition, high-resolution localization typically requires dense angular scans and, in narrowband implementations, repeated evaluations across frequency bins. These considerations motivate a localization strategy that combines broadband SH-domain processing, adequate angular resolution, and iterative multi-peak extraction with controlled computational complexity.

3. Direct sound and dominant reflection localization

The proposed localization procedure operates in the SH domain and estimates the DOAs of the direct sound and multiple dominant reflected components through an iterative DU peak-extraction scheme. It exploits the frequency-independent nature of SH steering vectors, which enables coherent broadband processing within the adopted dominant-path formulation for single-source reverberant environments.

3.1. SH-domain covariance estimation and frequency smoothing

Assuming quasi-stationarity over B snapshots, the SH-domain covariance matrix is estimated for each frequency bin f as

$$\widehat{\mathbf{R}}_{\mathrm{SH}}(f) = \frac{1}{B}\sum_{k=1}^{B} \mathbf{b}(k, f)\mathbf{b}^{\mathrm{H}}(k, f), \tag{15}$$

where $\mathbf{b}(k, f) \in \mathbb{C}^{(L+1)^2}$ denotes the SH-domain observation at time frame k and frequency bin f.

Exploiting the frequency-independent SH steering vectors, a single broadband covariance matrix is obtained via frequency smoothing with per-bin trace normalization (frequency smoothing power transform, FSPT) [25]:

$$\widehat{\mathbf{R}}_{\mathrm{FSPT}} = \sum_{f=f_{\min}}^{f_{\max}} \frac{\widehat{\mathbf{R}}_{\mathrm{SH}}(f)}{\mathrm{tr}\Big\{\widehat{\mathbf{R}}_{\mathrm{SH}}(f)\Big\}}. \tag{16}$$

This broadband fusion equalizes the contribution of individual frequency bins by normalizing each narrowband covariance to unit trace. Any additional global scaling, such as an explicit $1/F$ factor with $F = f_{\max} - f_{\min} + 1$, does not affect peak locations and is therefore omitted. Relative to an unnormalized sum, trace normalization reduces the influence of atypically strong bins and yields a better balanced broadband covariance representation.

3.2. Diagonal-unloading pseudo-spectrum

Let $\widehat{\mathbf{R}}^{(i)}$ denote the covariance matrix used at iteration i, initialized as

$$\widehat{\mathbf{R}}^{(0)} = \widehat{\mathbf{R}}_{\mathrm{FSPT}}. \tag{17}$$

At iteration i, diagonal unloading is performed as

$$\mathbf{R}_{\mathrm{DU}}^{(i)} = \hat{\lambda}_{\max}^{(i)}\,\mathbf{I} - \hat{\mathbf{R}}^{(i)}, \tag{18}$$

where $\hat{\lambda}_{\max}^{(i)}$ is the largest eigenvalue of $\hat{\mathbf{R}}^{(i)}$, estimated via the power method [35]. The resulting matrix defines a high-resolution DU pseudo-spectrum [22]. The DU pseudo-spectrum is evaluated on a discrete grid of candidate directions $\boldsymbol{\Omega} = [\theta, \vartheta]^{\top}$ as

$$\mathcal{P}_{\mathrm{DU}}^{(i)}(\boldsymbol{\Omega}) = \frac{1}{\mathbf{y}^{\mathrm{H}}(\boldsymbol{\Omega})\,\mathbf{R}_{\mathrm{DU}}^{(i)}\,\mathbf{y}(\boldsymbol{\Omega})}, \tag{19}$$

where $\mathbf{y}(\boldsymbol{\Omega}) \in \mathbb{C}^{(L+1)^2}$ is the SH steering vector. The direction associated with the dominant component at iteration i is obtained by grid search,

$$\hat{\boldsymbol{\Omega}}_i = \arg\max_{\boldsymbol{\Omega}}\ \mathcal{P}_{\mathrm{DU}}^{(i)}(\boldsymbol{\Omega}). \tag{20}$$

3.3. Iterative null-steering covariance update for coherent component extraction

To extract multiple coherent components, i.e., the direct sound and dominant reflections, the contribution associated with each previously estimated direction is suppressed through an iterative null-steering covariance update. The update relies on left–right covariance filtering rather than classical rank-one deflation based on estimated source powers or eigenvectors. Unlike rank-one subtraction, which depends directly on the accuracy of the component model, the proposed update suppresses the extracted direction through covariance-domain null steering and is therefore more stable under coherent multipath and successive extraction.

Let

$$\mathbf{d}_i = \mathbf{y}(\hat{\boldsymbol{\Omega}}_i), \qquad \alpha_i = \mathbf{d}_i^{\mathrm{H}}\mathbf{d}_i. \tag{21}$$

The null-steering (suppression) matrix is defined as

$$\mathbf{F}_i = \alpha_i\,\mathbf{I} - \mathbf{d}_i\mathbf{d}_i^{\mathrm{H}}. \tag{22}$$

By construction, $\mathbf{F}_i\mathbf{d}_i = \mathbf{0}$, so a spatial null is imposed along the extracted steering direction. This suppression step is related to null-steering beamforming and CLEAN-like iterative suppression strategies [26,27], but here it is applied directly to the broadband SH-domain covariance matrix.

The covariance matrix is first updated as

$$\tilde{\mathbf{R}}^{(i+1)} = \mathbf{F}_i\,\hat{\mathbf{R}}^{(i)}\,\mathbf{F}_i^{\mathrm{H}}, \tag{23}$$

and subsequently normalized to unit trace,

$$\hat{\mathbf{R}}^{(i+1)} = \frac{\tilde{\mathbf{R}}^{(i+1)}}{\mathrm{tr}\left\{\tilde{\mathbf{R}}^{(i+1)}\right\}}. \tag{24}$$

This per-iteration renormalization differs from the per-bin trace normalization used in (16). Its purpose is to keep the updated covariance on a comparable scale across iterations as progressively stronger components are suppressed.

The steps in (18)–(24) are repeated for a prescribed number of components. The extracted components are indexed according to the iteration order, with $i = 0$ denoting the first extracted dominant direction and the following iterations providing P additional dominant components. The resulting set of estimates is

$$\hat{\mathcal{S}} = \{\hat{\boldsymbol{\Omega}}_0, \hat{\boldsymbol{\Omega}}_1, \ldots, \hat{\boldsymbol{\Omega}}_P\}, \tag{25}$$

where $\hat{\boldsymbol{\Omega}}_0$ denotes the first extracted dominant direction, while $\hat{\boldsymbol{\Omega}}_p$, $p > 0$, denote the subsequently extracted dominant components ordered by the iterative extraction sequence. In the simulated and measured conditions considered in this study, the first extracted dominant direction is expected to correspond to the direct path.

Algorithm 1 summarizes the complete procedure.

Algorithm 1 SH-domain broadband iterative DU beamforming (direct path + dominant components).

STFT: compute $\mathbf{x}(k,f) = [x_1(k,f), \ldots, x_Q(k,f)]^{\top}$ for all k and f
SH encoding: for each (k,f), compute $\mathbf{b}(k,f) = \mathbf{E}_{L,\Lambda}(f)\mathbf{x}(k,f)$
Covariance matrices: for each f, compute $\hat{\mathbf{R}}_{\mathrm{SH}}(f) = \frac{1}{B}\sum_{k=1}^{B}\mathbf{b}(k,f)\mathbf{b}^{\mathrm{H}}(k,f)$
$\hat{\mathbf{R}}_{\mathrm{FSPT}} \leftarrow \sum_{f=f_{\min}}^{f_{\max}} \hat{\mathbf{R}}_{\mathrm{SH}}(f) / \mathrm{tr}\{\hat{\mathbf{R}}_{\mathrm{SH}}(f)\}$
$\hat{\mathbf{R}}^{(0)} \leftarrow \hat{\mathbf{R}}_{\mathrm{FSPT}}$
for $i = 0$ **to** P **do**
  $\hat{\lambda}_{\max}^{(i)} \leftarrow \textsc{PowerMethod}\left(\hat{\mathbf{R}}^{(i)}\right)$
  $\mathbf{R}_{\mathrm{DU}}^{(i)} \leftarrow \hat{\lambda}_{\max}^{(i)}\mathbf{I} - \hat{\mathbf{R}}^{(i)}$
  $\mathcal{P}_{\mathrm{DU}}^{(i)}(\boldsymbol{\Omega}) \leftarrow \left(\mathbf{y}^{\mathrm{H}}(\boldsymbol{\Omega})\mathbf{R}_{\mathrm{DU}}^{(i)}\mathbf{y}(\boldsymbol{\Omega})\right)^{-1}$
  $\hat{\boldsymbol{\Omega}}_i \leftarrow \arg\max_{\boldsymbol{\Omega}}\ \mathcal{P}_{\mathrm{DU}}^{(i)}(\boldsymbol{\Omega})$
  if $i < P$ **then**
    $\mathbf{d}_i \leftarrow \mathbf{y}(\hat{\boldsymbol{\Omega}}_i)$
    $\alpha_i \leftarrow \mathbf{d}_i^{\mathrm{H}}\mathbf{d}_i$
    $\mathbf{F}_i \leftarrow \alpha_i\mathbf{I} - \mathbf{d}_i\mathbf{d}_i^{\mathrm{H}}$
    $\tilde{\mathbf{R}}^{(i+1)} \leftarrow \mathbf{F}_i\,\hat{\mathbf{R}}^{(i)}\,\mathbf{F}_i^{\mathrm{H}}$
    $\hat{\mathbf{R}}^{(i+1)} \leftarrow \tilde{\mathbf{R}}^{(i+1)} / \mathrm{tr}\{\tilde{\mathbf{R}}^{(i+1)}\}$
  end if
end for
Output: $\hat{\mathcal{S}} = \{\hat{\boldsymbol{\Omega}}_0, \hat{\boldsymbol{\Omega}}_1, \ldots, \hat{\boldsymbol{\Omega}}_P\}$

3.4. Computational complexity and comparison with baselines

To compare computational cost with that of the baselines, let $K = P + 1$ denote the number of extracted components, i.e., the direct path plus P dominant reflected components. Let $G = N_{\mathrm{azim}}N_{\mathrm{col}}$ denote the number of candidate directions in the angular search grid.

The following expressions capture the dominant asymptotic scaling of the adopted implementations. They retain the leading-order cost of the main localization stages and omit lower-order constants, memory-access effects, and the STFT front end common to all methods.

For the proposed SH-domain iterative DU method, the main stages are SH encoding of all time–frequency observations, narrowband SH covariance accumulation, broadband covariance fusion via FSPT, iterative DU map formation and grid search over a single broadband covariance matrix, and per-iteration covariance update with trace renormalization. These stages scale as $\mathcal{O}(BFQN_{\mathrm{SH}})$, $\mathcal{O}(BFN_{\mathrm{SH}}^2)$, $\mathcal{O}(FN_{\mathrm{SH}}^2)$, $\mathcal{O}(KGN_{\mathrm{SH}}^2)$, and $\mathcal{O}(KN_{\mathrm{SH}}^3)$, respectively. The cubic term arises from the two-sided covariance update $\mathbf{F}_i\hat{\mathbf{R}}^{(i)}\mathbf{F}_i^{\mathrm{H}}$ when evaluated as a dense matrix–matrix product, whereas the subsequent trace renormalization alone is only quadratic in N_{SH}. The cost of estimating the dominant eigenvalue in (18) by a fixed-number power iteration is of lower order and is therefore absorbed into the dominant scaling.

Accordingly, the overall cost of the proposed method can be summarized as

$$\mathcal{C}_{\mathrm{Proposed}} = \mathcal{O}\left(BFQN_{\mathrm{SH}} + BFN_{\mathrm{SH}}^2 + FN_{\mathrm{SH}}^2 + KGN_{\mathrm{SH}}^2 + KN_{\mathrm{SH}}^3\right). \tag{26}$$

For the non-iterative SH-DU baseline, the same SH-domain preprocessing and broadband covariance construction are required, but only one DU map is evaluated. Its dominant cost is therefore

$$\mathcal{C}_{\mathrm{SH\text{-}DU}} = \mathcal{O}\left(BFQN_{\mathrm{SH}} + BFN_{\mathrm{SH}}^2 + FN_{\mathrm{SH}}^2 + GN_{\mathrm{SH}}^2\right). \tag{27}$$

For the SH-MUSIC baseline, a standard frequency-by-frequency implementation is considered. After SH encoding and narrowband covariance estimation, MUSIC requires an eigendecomposition at each analyzed bin and pseudo-spectrum evaluation over the angular grid at each

Table 1
Dominant computational complexity of the proposed method and the considered baselines.

Method	Dominant complexity
Proposed	$\mathcal{O}\left(BFQN_{\mathrm{SH}} + BFN_{\mathrm{SH}}^2 + FN_{\mathrm{SH}}^2 + KGN_{\mathrm{SH}}^2 + KN_{\mathrm{SH}}^3\right)$
SH-DU	$\mathcal{O}\left(BFQN_{\mathrm{SH}} + BFN_{\mathrm{SH}}^2 + FN_{\mathrm{SH}}^2 + GN_{\mathrm{SH}}^2\right)$
SH-MUSIC	$\mathcal{O}\left(BFQN_{\mathrm{SH}} + BFN_{\mathrm{SH}}^2 + FN_{\mathrm{SH}}^3 + FGN_{\mathrm{SH}}^2\right)$
SH-SRP	$\mathcal{O}\left(BFQN_{\mathrm{SH}} + BFN_{\mathrm{SH}}^2 + FGN_{\mathrm{SH}}^2\right)$
SRP-PHAT	$\mathcal{O}\left(BFQ^2 + FQ^2 + FGQ^2\right)$

frequency. Its dominant cost can therefore be expressed as

$$C_{\text{SH-MUSIC}} = \mathcal{O}\left(BFQN_{\mathrm{SH}} + BFN_{\mathrm{SH}}^2 + FN_{\mathrm{SH}}^3 + FGN_{\mathrm{SH}}^2\right). \tag{28}$$

The cubic dependence on N_{SH} arises from the per-frequency eigendecomposition, whereas the grid-search stage remains quadratic in N_{SH} for each evaluated direction.

For the SH-SRP baseline, again using a standard frequency-by-frequency implementation, the dominant cost is the SH-domain grid evaluation at each analyzed frequency bin, yielding

$$C_{\text{SH-SRP}} = \mathcal{O}\left(BFQN_{\mathrm{SH}} + BFN_{\mathrm{SH}}^2 + FGN_{\mathrm{SH}}^2\right). \tag{29}$$

For the microphone-domain SRP-PHAT baseline, a covariance-matrix implementation is considered. In particular, the method is implemented through PHAT-normalized microphone-domain covariance matrices. Its dominant operations are therefore microphone-domain covariance estimation, PHAT normalization, and grid evaluation at each analyzed frequency bin. The corresponding dominant cost can be summarized as

$$C_{\text{SRP-PHAT}} = \mathcal{O}\left(BFQ^2 + FQ^2 + FGQ^2\right). \tag{30}$$

These expressions clarify the main differences among the considered methods. Relative to SH-DU, the proposed method adds iterative covariance updates and repeated broadband map evaluations, with cost roughly proportional to the number of extracted components K. By contrast, SH-MUSIC, SH-SRP, and the considered microphone-domain SRP-PHAT implementation use frequency-by-frequency processing and therefore retain an explicit dependence on F in the grid-search stage. Among the SH-domain baselines, SH-DU is the most favorable in complexity because it shares the same broadband covariance construction as the proposed method while avoiding both iterative suppression and per-frequency eigendecomposition, whereas SH-MUSIC is the most demanding because of its per-frequency cubic eigendecomposition cost. The proposed method therefore occupies an intermediate position: it is lighter than frequency-by-frequency MUSIC but more expensive than SH-DU because of the iterative covariance-update stage.

Table 1 summarizes the dominant asymptotic complexity of the proposed method and the considered baselines.

4. Simulations

The simulation study has three parts. First, the proposed method is evaluated under a fixed physical array configuration to isolate the effect of the retained SH order. Second, it is evaluated under order-dependent configurations, in which both the number of microphones and the array radius vary with SH order under a minimum-spacing constraint. Third, it is compared with four baseline methods under a common high-order reverberant condition.

All simulations are conducted in a rectangular reverberant room using the image-source method (ISM) [36]. A single active acoustic source is assumed, and the simulated sound field at the array consists of the direct path and multiple specular reflections generated by the room boundaries.

4.1. Room geometry, source positions, and image-source references

The simulated room dimensions are $\mathbf{D}_{\mathrm{room}} = [7,\ 9,\ 4]$ m, and the array center is fixed at $\mathbf{r}_c = [2.2,\ 4.7,\ 1.5]$ m. For the proposed-method analyses, two reverberation conditions are considered, namely $\mathrm{T}_{60} = 0.75$ s and $\mathrm{T}_{60} = 1.5$ s. In addition, the baseline comparison is carried out at $\mathrm{T}_{60} = 1.0$ s using the highest SH order considered in the simulations, namely $L = 15$, together with the corresponding largest physical array configuration.

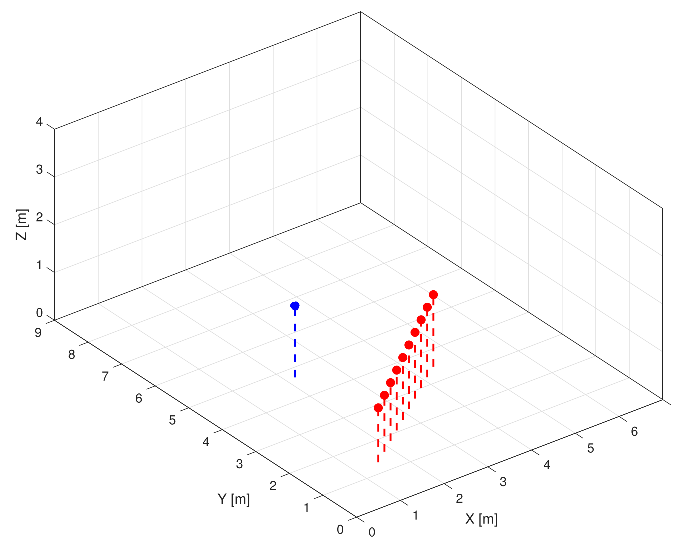

Fig. 1. Simulated room configuration showing the array and source positions.

The source position is varied across $T = 10$ trials, as shown in Fig. 1. For each source position, room impulse responses are generated with the ISM. Ground-truth candidate directions for the reflected components are obtained from an image-source construction. The candidate pool used for greedy matching is built by generating all image sources up to third order, that is, all integer triplets (n_x, n_y, n_z) such that $|n_x| + |n_y| + |n_z| \le 3$, excluding the direct source and the six axis-aligned third-order terms $(\pm 3, 0, 0)$, $(0, \pm 3, 0)$, and $(0, 0, \pm 3)$. This yields $N_{\mathrm{img}} = 56$ candidate image sources for reflection matching in the error computation.

4.2. Array configurations

The array consists of Q omnidirectional microphones approximately uniformly distributed on a sphere according to a Fibonacci lattice [37]. The spherical coordinates of the qth microphone on the unit sphere are

$$\vartheta_q = \arccos\left(1 - \frac{2(q - 0.5)}{Q}\right), \tag{31}$$

$$\theta_q = \mathrm{mod}\left(\pi(1 + \sqrt{5})(q - 0.5),\ 2\pi\right), \tag{32}$$

where $\theta_q \in [0, 2\pi)$ denotes the azimuth and $\vartheta_q \in [0, \pi]$ denotes the colatitude. The associated Cartesian coordinates are

$$\mathbf{r}_q^{(\mathrm{unit})} = \begin{bmatrix} \sin\vartheta_q \cos\theta_q \\ \sin\vartheta_q \sin\theta_q \\ \cos\vartheta_q \end{bmatrix}. \tag{33}$$

Four SH orders are considered, namely $L \in \{1, 5, 10, 15\}$. The corresponding theoretical minimum microphone counts satisfying $(L+1)^2 \le Q$ are

$$Q_{\min} \in \{4, 36, 121, 256\}. \tag{34}$$

Two array-configuration studies are considered. In the first, referred to as the fixed physical configuration, the same maximum physical array is used for all SH orders. Specifically, the array always employs $Q = 256$ microphones, and the radius is kept equal to that of the highest-order configuration. The retained SH analysis order is then varied as $L \in \{1, 5, 10, 15\}$, while the physical aperture and the number of microphones remain unchanged. This setting isolates the effect of the retained SH order from changes in the physical array.

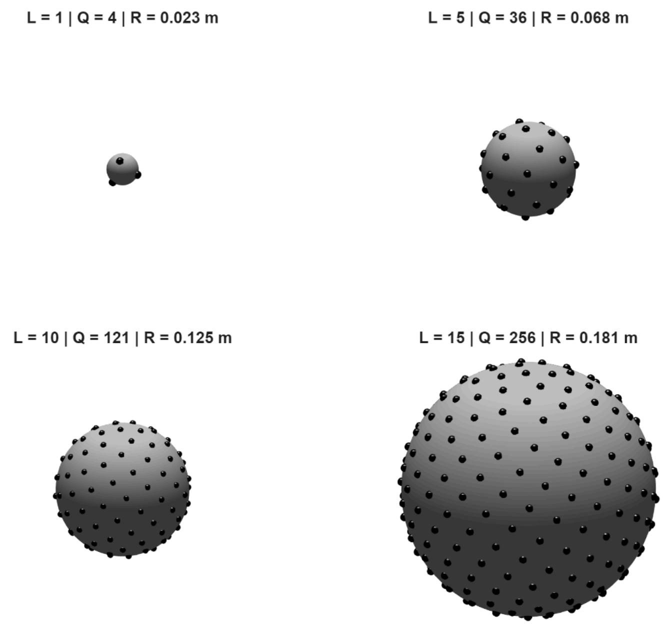

Fig. 2. The order-dependent physical SMA configurations used in the simulations for the different microphone counts Q.

In the second, referred to as the order-dependent physical configuration, each SH order L is associated with its minimum microphone count, namely $Q_{\min} = (L+1)^2$. For each such configuration, the array radius is selected by imposing a minimum inter-microphone spacing $d_{\min} = 0.035$ m, namely

$$R = \frac{d_{\min}}{\min\limits_{q \neq q'} \left\| \mathbf{r}_q^{(\text{unit})} - \mathbf{r}_{q'}^{(\text{unit})} \right\|}, \tag{35}$$

where the minimum is computed over the unit-sphere microphone layout corresponding to the selected value of Q. In this case, both the number of microphones and the physical aperture increase with L. This setting therefore captures the joint effect of increasing SH order and physically scaling the array under a realistic spacing constraint. The array geometries corresponding to the considered SH orders are illustrated in Fig. 2.

4.3. Signal generation and perturbation model

The source signal is a speech recording sampled to $f_s = 24$ kHz and processed in the STFT domain using a Hann window of length $N_w = 1024$ samples and a hop size of $N_w/4$ (75% overlap). The analyzed frequency band is [80, 8000] Hz. SH encoding follows (10)–(11), with adaptive regularization scaling $\beta_{\text{set}} = 5 \times 10^{-3}$. The choice of β_{set} remains an implementation parameter of the regularized SH encoding stage. In this study, it is selected to provide stable broadband processing for the considered arrays and frequency range, but it is not claimed to be universally optimal across all array configurations and acoustic conditions.

Additive spatially white sensor noise is introduced according to a target SNR. In addition, an isotropic diffuse-noise field is added and is controlled through the SDNR [38]. Finally, microphone gain mismatch can be introduced by applying independent random gains to the channels, parametrized by the standard deviation of the gain perturbation in dB.

For the two proposed-method studies, namely the fixed physical configuration and the order-dependent physical configuration, the reference acoustic condition is SNR = 30 dB, SDNR = 30 dB, and $\sigma_{\text{gain}} = 0$ dB.

For the baseline comparison at $T_{60} = 1.0$ s, four acoustic conditions are considered:

$$C_1 : \text{SNR} = 30\ \text{dB},\ \text{SDNR} = 30\ \text{dB},\ \sigma_{\text{gain}} = 0\ \text{dB}, \tag{36}$$

$$C_2 : \text{SNR} = 10\ \text{dB},\ \text{SDNR} = 30\ \text{dB},\ \sigma_{\text{gain}} = 0\ \text{dB}, \tag{37}$$

$$C_3 : \text{SNR} = 30\ \text{dB},\ \text{SDNR} = 10\ \text{dB},\ \sigma_{\text{gain}} = 0\ \text{dB}, \tag{38}$$

$$C_4 : \text{SNR} = 10\ \text{dB},\ \text{SDNR} = 10\ \text{dB},\ \sigma_{\text{gain}} = 1\ \text{dB}. \tag{39}$$

4.4. Angular grid, extracted components, and evaluation metrics

Directional searches are performed on a uniform angular grid with 1° resolution in azimuth and colatitude. In the simulations, the direct path plus 15 additional dominant components are evaluated, for a total of $K = 16$ extracted directions.

Given a direction $\boldsymbol{\Omega} = [\theta, \vartheta]^\top$, the associated unit vector is

$$\mathbf{u}(\theta, \vartheta) = \begin{bmatrix} \sin\vartheta \cos\theta \\ \sin\vartheta \sin\theta \\ \cos\vartheta \end{bmatrix}. \tag{40}$$

For each trial, the estimated directions are matched to the ISM candidate pool using greedy nearest-neighbor association in angular distance. After each match, the selected candidate is removed from the available set. This protocol is used for all methods considered.

Given an estimate $\hat{\boldsymbol{\Omega}} = [\hat{\theta}, \hat{\vartheta}]^\top$ and a reference $\boldsymbol{\Omega} = [\theta, \vartheta]^\top$, the angular error is computed as

$$d(\hat{\boldsymbol{\Omega}}, \boldsymbol{\Omega}) = \frac{180}{\pi} \arccos\Big(\text{clip}\left(\mathbf{u}^\top(\hat{\theta}, \hat{\vartheta}) \mathbf{u}(\theta, \vartheta), -1, 1 \right) \Big), \tag{41}$$

where $\text{clip}(x, a, b) = \min(\max(x, a), b)$ limits its argument to the interval $[a, b]$ for numerical stability.

Two aggregate metrics are reported for each experiment: one computed over the direct path and the first six extracted dominant components (D + R1–R6), and one computed over all extracted components (D + R1–R15). Let $\mathcal{P}_6 = \{0, 1, \dots, 6\}$ and $\mathcal{P}_{15} = \{0, 1, \dots, 15\}$ denote the corresponding index sets. The aggregate RMSE over a generic subset $\mathcal{P}$ is defined as

$$\text{RMSE}_{\mathcal{P}} = \sqrt{\frac{1}{T\,|\mathcal{P}|} \sum_{p \in \mathcal{P}} \sum_{t=1}^{T} \left(d\left(\hat{\boldsymbol{\Omega}}_p^{(t)}, \boldsymbol{\Omega}_p^{(t)}\right) \right)^2}, \tag{42}$$

and the corresponding aggregate AR is defined as

$$\text{AR}_{\mathcal{P}} = \frac{1}{T\,|\mathcal{P}|} \sum_{p \in \mathcal{P}} \sum_{t=1}^{T} \mathbb{I}\Big(d\left(\hat{\boldsymbol{\Omega}}_p^{(t)}, \boldsymbol{\Omega}_p^{(t)}\right) \leq \tau_{\text{AR}} \Big), \tag{43}$$

where $\mathbb{I}(\cdot)$ is the indicator function. Throughout this study, the threshold is set to $\tau_{\text{AR}} = 5°$. Accordingly, the table entries denoted as RMSE (D + R1–R6) and AR (D + R1–R6) correspond to (42) and (43) evaluated over $\mathcal{P}_6$, whereas RMSE (All) and AR (All) correspond to the same quantities evaluated over $\mathcal{P}_{15}$.

4.5. Proposed method with fixed physical configuration

Table 2 reports the aggregate metrics for the fixed-configuration study. Because the physical array is unchanged across all cases, the results isolate the effect of the retained SH analysis order on multi-component localization.

Performance improves markedly as the retained SH order increases. For both reverberation times, very low order is insufficient for reliable multi-component localization, whereas intermediate order already yields a substantial improvement. High-order processing achieves near-perfect AR for the direct path and the first six dominant reflected components, together with much lower RMSE when all extracted components are included. These results indicate that, under fixed aperture and microphone count, the retained SH order is the main factor enabling accurate localization of the direct path and the strongest dominant reflected components.

Table 2
Proposed method for the fixed physical configuration study: the same maximum array configuration is used for all SH orders, namely $Q = 256$ microphones and the corresponding fixed radius. The acoustic condition is SNR = 30 dB, SDNR = 30 dB, and $\sigma_{\text{gain}} = 0$ dB.

T_{60} [s]	L	RMSE (D + R1–R6)	RMSE (All)	AR (D + R1–R6)	AR (All)
0.75	1	19.125	20.188	0.16	0.12
	5	3.833	6.770	0.80	0.56
	10	1.500	2.252	1.00	0.99
	15	0.843	1.829	1.00	0.95
1.5	1	18.267	20.085	0.09	0.09
	5	4.167	7.341	0.74	0.50
	10	1.956	2.692	1.00	0.98
	15	1.075	2.171	1.00	0.94

Table 3
Proposed method for the order-dependent physical configuration study: for each SH order L, the minimum microphone count $Q = (L + 1)^2$ is used, and the corresponding radius is selected according to the minimum-spacing constraint. The acoustic condition is SNR = 30 dB, SDNR = 30 dB, and $\sigma_{\text{gain}} = 0$ dB.

T_{60} [s]	L	RMSE (D + R1–R6)	RMSE (All)	AR (D + R1–R6)	AR (All)
0.75	1	18.322	18.779	0.09	0.08
	5	5.816	9.908	0.71	0.46
	10	1.671	2.637	1.00	0.97
	15	0.843	1.829	1.00	0.95
1.5	1	19.250	19.373	0.10	0.09
	5	5.181	9.562	0.64	0.41
	10	2.227	3.057	1.00	0.95
	15	1.075	2.171	1.00	0.94

4.6. Proposed method with order-dependent physical configuration

Table 3 reports the aggregate metrics for the order-dependent configuration study. In this case, the results reflect the joint effect of increasing SH order, microphone count, and physical aperture under the adopted minimum-spacing constraint.

The same overall trend remains: performance improves substantially as the SH order increases, and the most reliable localization is obtained at high order. Here, however, the gain cannot be attributed to the retained SH order alone, because the number of microphones and the array aperture also increase with L. The results therefore confirm the benefit of the proposed method under a more realistic scaling scenario, in which higher-order processing is accompanied by denser spatial sampling and larger physical aperture.

4.7. Comparison with baseline methods

The baseline comparison is conducted under a common high-order reverberant configuration. The compared methods are the proposed iterative SH-DU approach, SH-DU, SH-MUSIC, SH-SRP, and microphone-domain SRP-PHAT. To enable a controlled multi-peak comparison with the proposed iterative extraction, $K = 16$ directions were selected from the broadband map using spherical non-maximum suppression with an exclusion radius $\Delta = 8°$. Specifically, peaks were extracted sequentially by repeatedly selecting the current global maximum and suppressing all grid points within an angular distance Δ from the selected peak. Azimuth wrap-around was applied in the suppression step, i.e., neighborhoods crossing $0°/360°$ were handled circularly in azimuth.

Table 4 reports aggregate localization results under the four acoustic conditions. For the proposed method, SH-DU, SH-MUSIC, and SH-SRP, the analyzed frequency range is 80–8000 Hz, consistent with the main simulation setup. For SRP-PHAT, the range is restricted to 200–4000 Hz to keep the microphone-domain implementation computationally feasible for the adopted array size and angular resolution. This restriction is due to implementation constraints: in the microphone domain, steering vectors are frequency-dependent, and storing them over the full analyzed band with the adopted high-order array and dense angular grid would require excessive memory.

Table 4
Comparison with baseline methods at $T_{60} = 1.0$ s, $L = 15$, and the maximum array configuration. Condition C_1: (30, 30, 0) dB; condition C_2: (10, 30, 0) dB; condition C_3: (30, 10, 0) dB; condition C_4: (10, 10, 1) dB, where the triplets denote (SNR, SDNR, σ_{gain}).

Cond.	Method	RMSE (D + R1–R6)	RMSE (All)	AR (D + R1–R6)	AR (All)
C_1	Proposed	0.898	2.011	1.00	0.93
	SH-DU	9.492	14.489	0.51	0.43
	SH-MUSIC	9.991	15.312	0.51	0.41
	SH-SRP	10.145	16.837	0.43	0.29
	SRP-PHAT	12.643	22.189	0.24	0.11
C_2	Proposed	3.291	8.384	0.94	0.71
	SH-DU	9.092	11.904	0.47	0.47
	SH-MUSIC	9.519	14.336	0.50	0.43
	SH-SRP	9.967	15.624	0.44	0.33
	SRP-PHAT	13.335	22.856	0.23	0.10
C_3	Proposed	0.963	3.976	1.00	0.88
	SH-DU	9.413	14.177	0.50	0.44
	SH-MUSIC	9.920	15.160	0.50	0.41
	SH-SRP	10.145	16.898	0.43	0.28
	SRP-PHAT	13.464	23.235	0.24	0.11
C_4	Proposed	4.841	8.709	0.87	0.71
	SH-DU	9.466	11.528	0.47	0.49
	SH-MUSIC	8.695	13.520	0.51	0.42
	SH-SRP	10.104	13.895	0.39	0.28
	SRP-PHAT	13.350	23.209	0.24	0.11

Table 5
Runtime comparison at $L = 15$, $Q = 256$, maximum array radius, and angular grid identical to that used in Table 4. The reported times are averaged over conditions C_1–C_4. Runtime is reported relative to the proposed method.

Method	Average time [s]	Runtime ratio
Proposed	172.81	1.00×
SH-DU	171.01	0.99×
SH-MUSIC	270.67	1.57×
SH-SRP	268.77	1.56×
SRP-PHAT	281.07	1.63×

The proposed method yields the best localization performance in all four acoustic conditions and for both aggregate metrics. The advantage is apparent for the direct path and strongest dominant reflected components and becomes even clearer when all extracted components are included. Among the baselines, SH-DU is generally the strongest overall competitor, especially over the full extraction set, whereas SH-MUSIC remains broadly competitive on the D + R1–R6 subset. By contrast, SH-SRP performs consistently worse than SH-DU and SH-MUSIC, and SRP-PHAT shows the weakest overall localization performance, particularly when evaluation is extended to all extracted components. Overall, the results indicate that iterative covariance suppression improves not only the first extraction steps but also performance at greater extraction depth under adverse acoustic conditions.

The runtime comparison is reported separately in Table 5. The reported times are obtained on the same 2.675 s speech signal and include the full processing chain, from windowing and STFT analysis to SH-domain encoding, covariance estimation, acoustic-map evaluation, and the corresponding extraction stages. The proposed method has a runtime very close to that of SH-DU, indicating that the additional iterative covariance-update stage introduces only limited overhead in the present implementation. SH-MUSIC, SH-SRP, and SRP-PHAT are all substantially slower. Overall, the results support the computational attractiveness of the proposed broadband iterative SH-domain formulation relative to the considered alternatives.

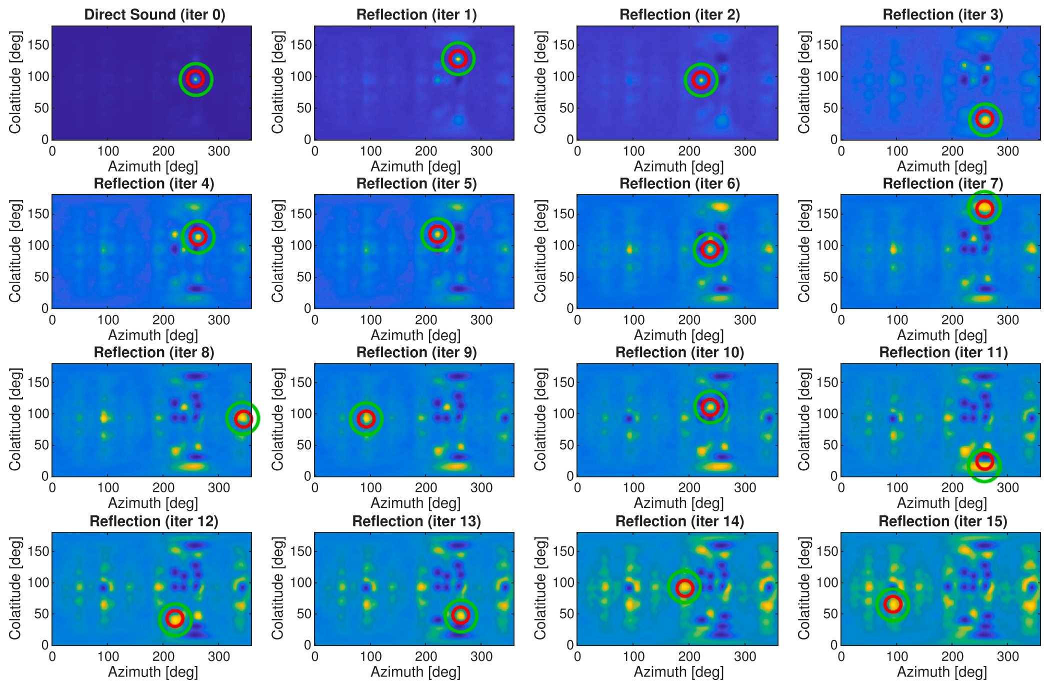

Fig. 3. Illustrative iterative DU acoustic maps obtained with the proposed method in a controlled simulated high-order configuration (L = 15) with T_{60} = 1.5 s, SNR = 30 dB, SDNR = 30 dB, and σ_{gain} = 0 dB. The displayed maps qualitatively illustrate the successive extraction mechanism under favorable resolution conditions, while the quantitative performance under both reference and more challenging acoustic conditions is reported in the simulation tables. Green marker: ground-truth DOA (matched ISM candidate). Red marker: estimated DOA.

4.8. Discussion of simulation results

Taken together, the simulation results show that higher-order SH processing is the main enabler of accurate multi-component localization. Under the fixed physical configuration, where aperture and microphone count are constant, the improvement with increasing L can be attributed primarily to the higher angular resolution provided by higher-order SH analysis. The order-dependent configuration confirms the same trend in a more realistic setting, where the gains reflect the combined effect of higher SH order, denser spatial sampling, and larger aperture.

The baseline comparison further shows that iterative covariance suppression is the main factor behind the robustness of the proposed method. Relative to the non-iterative baselines, the proposed approach consistently yields lower RMSE and higher AR, with the clearest advantage appearing when evaluation is extended to the full extraction set. The comparison with SH-DU is particularly informative because SH-DU uses the same SH-domain broadband covariance construction but does not include iterative suppression. The clear performance gap between the two methods, together with their very similar runtime in the present implementation, suggests that the covariance-update stage offers a favorable tradeoff between extraction capability and computational cost.

The simulations also confirm that deep extraction remains intrinsically more difficult than localization of the direct path and the strongest dominant reflected components. Even for the proposed method, performance degrades when evaluation is extended from D + R1–R6 to the full extraction set, especially under more adverse acoustic conditions. The iterative formulation should therefore be viewed not as eliminating the difficulty of coherent multipath extraction, but as improving its stability and practical effectiveness relative to the considered baselines.

To complement the aggregate RMSE and AR results, Fig. 3 provides a qualitative illustration of the successive extraction process in a controlled high-order simulated configuration. The corresponding quantitative trends are reported in Tables 2–5.

4.9. Summary of simulation findings and limitations

Overall, the simulations show that higher SH order is critical for accurate multi-component localization and that the proposed iterative SH-domain DU method consistently outperforms the considered baselines, especially at greater extraction depth.

At the same time, the simulation framework should be interpreted as a controlled basis for algorithmic comparison. It does not explicitly model effects such as non-specular scattering, frequency-dependent wall absorption, or full calibration errors beyond the gain-mismatch perturbation considered here. This limitation motivates the measured-data experiments in the next section, which are used to verify whether the main simulation trends remain visible under real-room conditions.

5. Measured-data experiments

To complement the controlled simulation study, additional experiments are conducted on measured room-acoustic data from the public Arni spatial room dataset [28]. The analysis uses real multichannel room impulse responses recorded with a 32-channel mh Acoustics Eigenmike EM32 SMA. The experiments reported here use the SOFA file corresponding to source position S1 in the reverberant condition. The measured responses are provided at 48 kHz together with metadata describing the source and array poses. Because the dataset documentation does not report a reference T_{60} value for the measured room, the reverberation time is estimated directly from the measured room impulse responses in octave bands according to ISO 3382 using a noise-compensated decay analysis [39]. The estimates are highly consistent across the analyzed responses, with an average T_{60} of approximately 0.81 s.

To keep the evaluation compact while retaining representative variability, the real-data analysis uses 10 measured receiver positions selected from `arni_em32_reverberant_S1.sofa`. The same general processing chain used in simulation is retained, including STFT analysis, SH-domain encoding, broadband covariance construction, and

Table 6
Measured-data comparison on 10 selected positions from `arni_em32_reverberant_S1.sofa` using the 32-channel Eigenmike recordings. The analysis is carried out at the maximum supported SH order $L = 4$.

Method	RMSE (D + R1–R6)	RMSE (All)	AR (D + R1–R6)	AR (All)
Proposed	6.757	12.667	0.61	0.36
SH-DU	9.119	20.127	0.49	0.21
SH-MUSIC	8.685	19.496	0.47	0.21
SH-SRP	9.005	20.469	0.51	0.23
SRP-PHAT	9.234	19.446	0.51	0.23

comparison with the same baselines. The measured impulse responses are first resampled from 48 kHz to the processing rate of 24 kHz, and the source signal is then convolved with the resampled multichannel responses. Because the EM32 array has $Q = 32$ microphones, the measured-data analysis is limited to a maximum SH order of $L = 4$, which is therefore adopted throughout this section. Accordingly, the processing frequency range is restricted to 200–4000 Hz, as the lower SH order yields a less accurate spatial representation at higher frequencies, where truncation effects become more pronounced. The regularization parameter in the SH encoding stage is set to $\beta_{\text{set}} = 5 \times 10^{-2}$. The larger value, relative to that used in the simulations, compensates for the lower spatial oversampling of the EM32 at reduced SH order.

For the measured-data evaluation, the reference directions are constructed analogously to the simulation protocol using the known room dimensions together with the source and receiver positions provided by the dataset metadata. For each selected receiver position, the direct-path DOA is obtained from the source-to-array-center vector, and the reflected-path candidate pool is obtained from an image-source construction in the same room geometry. As in the simulation study, all image sources up to third order are generated, the direct source and the six axis-aligned third-order terms are excluded, and greedy nearest-neighbor angular matching is used to associate the extracted directions with the reference candidate set. These geometry-based image-source directions should be interpreted as reference proxies for dominant early propagation components rather than exact ground truth for all extracted reflections in the real room.

Table 6 shows that the proposed method provides the best overall performance on the measured Eigenmike data. It achieves both the lowest RMSE and the highest AR at 5° for the D + R1–R6 aggregate and for the full extraction set. These results suggest that iterative covariance suppression improves the stability of multi-component extraction under real-room conditions as well. All baseline methods degrade markedly when evaluation is extended to the full set of extracted components, indicating that deeper extraction remains substantially more challenging without iterative suppression. Overall, the measured-data results are consistent with the main trend observed in simulation, namely that iterative SH-domain DU processing improves localization of the direct path and dominant reflected components relative to the considered non-iterative SH-domain and microphone-domain baselines. At the same time, absolute performance remains below that observed in the most favorable simulated conditions, as expected given the greater complexity of the measured sound field and the lower maximum SH order supported by the EM32 array.

6. Conclusions

This paper addressed localization of the direct path together with a relatively large set of dominant reflected components in coherent single-source reverberant environments using SMAs. To this end, a broadband iterative DU framework was developed in the SH domain. The framework combines frequency-smoothed covariance fusion, high-resolution DU map evaluation, and iterative null-steering covariance filtering. By operating on a single broadband SH covariance matrix, the proposed approach avoids narrowband per-frequency spatial searches during multi-peak extraction and retains a computational structure that remains attractive for high-order SH processing.

The method was evaluated through controlled simulations across SH orders, array configurations, and adverse acoustic conditions, and was further validated on measured multichannel room impulse responses. The results indicate that the proposed framework provides an effective and scalable strategy for structured localization of direct sound and dominant reflected components in reverberant rooms. In particular, increasing the SH order improves localization of the direct path and strongest dominant reflected components, whereas deeper extraction stages remain more challenging as weaker and more ambiguous components are targeted. The measured-data experiments further confirm the advantage of the iterative SH-domain DU strategy over the considered SH-domain and microphone-domain baselines.

From a methodological standpoint, the main strength of the proposed approach lies in its broadband SH-domain iterative formulation, which supports coherent multipath analysis without requiring computationally heavier frequency-by-frequency high-resolution pipelines. It therefore offers a favorable compromise among angular resolution, extraction capability, and computational cost, making the framework particularly relevant for high-order spherical-array processing, where narrowband subspace-based strategies can become increasingly expensive.

Overall, the proposed iterative SH-domain DU formulation provides an effective basis for multi-component DOA analysis in reverberant environments and for further developments in reflection-aware spatial audio and room-acoustics inference.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] M.J. Taghizadeh, A. Asaei, S. Haghighatshoar, P.N. Garner, H. Bourlard, Spatial sound localization via multipath euclidean distance matrix recovery, IEEE J. Sel. Top. Signal Process. 9 (5) (2015) 802–814.
[2] 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 (2017) 1–24.
[3] D. Salvati, C. Drioli, G.L. Foresti, Exploiting CNNs for improving acoustic source localization in noisy and reverberant conditions, IEEE Trans. Emerg. Top. Comput. Intell. 2 (2) (2018) 103–116.
[4] S. Lin, Reverberation-robust localization of speakers using distinct speech onsets and multichannel cross correlations, IEEE/ACM Trans. Audio Speech Lang. Process. 26 (11) (2018) 2098–2111.
[5] T. Long, J. Chen, G. Huang, J. Benesty, I. Cohen, Acoustic source localization based on geometric projection in reverberant and noisy environments, IEEE J. Sel. Top. Signal Process. 13 (1) (2019) 143–155.
[6] J. Yang, X. Zhong, W. Chen, W. Wang, Multiple acoustic source localization in microphone array networks, IEEE/ACM Trans. Audio Speech Lang. Process. 29 (2021) 334–347.
[7] M.J. Bianco, S. Gannot, E. Fernandez-Grande, P. Gerstoft, Semi-supervised source localization in reverberant environments with deep generative modeling, IEEE Access 9 (2021) 84956–84970.
[8] D. Salvati, C. Drioli, G.L. Foresti, Acoustic source localization using a geometrically sampled grid SRP-PHAT algorithm with max-pooling operation, IEEE Signal Process. Lett. 29 (2022) 1828–1832.
[9] J. Yan, W. Zhao, Y.I. Wu, Y. Zhou, Indoor sound source localization under reverberation by extracting the features of sample covariance, Appl. Acoust. 210 (2023) 109453.
[10] G. García-Barrios, E. Latorre Iglesias, J.M. Gutiérrez-Arriola, R. Fraile, N. Sáenz-Lechón, V.J. Osma-Ruiz, Exploiting spatial diversity for increasing the robustness of sound source localization systems against reverberation, Appl. Acoust. 202 (2023) 109138.
[11] W.-T. Lai, L. Birnie, T. Abhayapala, A. Bastine, P. Samarasinghe, Sound source localization using multi-dictionary orthogonal matching pursuit in reverberant environments, IEEE Trans. Audio Speech Lang. Process. 33 (2025) 4844–4855.

[12] F. Ribeiro, C. Zhang, D.A. Florencio, D.E. Ba, Using reverberation to improve range and elevation discrimination for small array sound source localization, IEEE Trans. Audio Speech Lang. Process. 18 (7) (2010) 1781–1792.
[13] F. Antonacci, J. Filos, M.R.P. Thomas, E.A.P. Habets, A. Sarti, P. Naylor, S. Tubaro, Inference of room geometry from acoustic impulse responses, IEEE Trans. Audio Speech Lang. Process. 20 (10) (2012) 2683–2695.
[14] F.H. de Araujo, F.A. de N. Castro Pinto, J.C. Boscher Torres, Room reflections analysis with the use of spherical beamforming and wavelets, Appl. Acoust. 131 (2018) 192–202.
[15] M. Lovedee-Turner, D. Murphy, Three-dimensional reflector localisation and room geometry estimation using a spherical microphone array, J. Acoust. Soc. Am. 146 (5) (2019) 3339–3352.
[16] W. Yu, W.B. Kleijn, Room acoustical parameter estimation from room impulse responses using deep neural networks, IEEE/ACM Trans. Audio Speech Lang. Process. 29 (2021) 436–447.
[17] S. Tervo, Single snapshot detection and estimation of reflections from room impulse responses in the spherical harmonic domain, IEEE/ACM Trans. Audio Speech Lang. Process. 24 (12) (2016) 2466–2480.
[18] E. Mabande, K. Kowalczyk, H. Sun, W. Kellermann, Room geometry inference based on spherical microphone array eigenbeam processing, J. Acoust. Soc. Am. 134 (4) (2013) 2773–2789.
[19] Y. Hadadi, H. Beit-On, V. Tourbabin, Z. Ben-Hur, D.L. Alon, B. Rafaely, Blind localization of early room reflections based on microphone arrays and reverberant speech, IEEE Trans. Audio Speech Lang. Process. 33 (2025) 3401–3413.
[20] J.-W. Yeowi, E.-L. Tan, J. Bai, S. Peksi, W.S. Gan, Enhancing 3-D sound event localization and detection with distance estimation using reverberation and spatial coherence features, IEEE Sens. J. 25 (15) (2025) 29221–29237.
[21] D. Salvati, C. Drioli, G.L. Foresti, A low-complexity robust beamforming using diagonal unloading for acoustic source localization, IEEE/ACM Trans. Audio Speech Lang. Process. 26 (3) (2018) 609–622.
[22] D. Salvati, C. Drioli, G.L. Foresti, Power method for robust diagonal unloading localization beamforming, IEEE Signal Process. Lett. 26 (5) (2019) 725–729.
[23] R.O. Schmidt, Multiple emitter location and signal parameter estimation, IEEE Trans. Antennas Propag. 34 (3) (1986) 276–280.
[24] J. Capon, High resolution frequency-wavenumber spectrum analysis, Proc. IEEE 57 (8) (1969) 1408–1418.
[25] D. Salvati, C. Drioli, G.L. Foresti, Diagonal unloading beamforming in the spherical harmonic domain for acoustic source localization in reverberant environments, IEEE/ACM Trans. Audio Speech Lang. Process. 28 (2020) 2001–2012.
[26] P. Sijtsma, CLEAN Based on spatial source coherence, Int. J. Aeroacoustics 6 (4) (2007) 357–374.
[27] D. Salvati, C. Drioli, G.L. Foresti, Iterative diagonal unloading beamforming for multiple acoustic sources localization using compact sensor arrays, IEEE Sens. J. 21 (13) (2021) 15080–15089.
[28] A.A. Figueroa Durán, N. Meyer-Kahlen, S.A. Verburg Riezu, K.T. Lokki, E. Fernandez Grande, Frequency and spatial limitations of sound field reconstruction methods for navigable reproduction, in: Forum Acusticum Euronoise 2025, Málaga, Spain, 2025.
[29] W. Wang, S. Yan, L. Mao, On the statistical performance of spherical harmonics MUSIC, Signal Process. 174 (2020) 107622.
[30] B. Rafaely, Spherical array beamforming, in: Fundamentals of Spherical Array Processing, 8 of *Springer Topics in Signal Processing*, Springer, 2015, pp. 101–123.
[31] J.H. DiBiase, H.F. Silverman, M.S. Brandstein, Microphone Arrays: Signal Processing Techniques and Applications, Springer, 2001.
[32] J. Meyer, G. Elko, A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2, 2002, pp. II–1781–II–1784.
[33] C.T. Jin, N. Epain, A. Parthy, Design, optimization and evaluation of a dual-radius spherical microphone array, IEEE/ACM Trans. Audio Speech Lang. Process. 22 (1) (2014) 193–204.
[34] P.C. Hansen, Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion, SIAM, 1998.
[35] G.H. Golub, C.F.V. Loan, Matrix Computations, Johns Hopkins University Press, 1996.
[36] J.B. Allen, D.A. Berkley, Image method for efficiently simulating small-room acoustics, J. Acoust. Soc. Am. 65 (4) (1979) 943–950.
[37] A. González, Measurement of areas on a sphere using Fibonacci and latitude–longitude lattices, Math. Geosci. 42 (1) (2010) 49–64.
[38] E.A.P. Habets, S. Gannot, Generating sensor signals in isotropic noise fields, J. Acoust. Soc. Am. 122 (6) (2007) 3464–3470.
[39] International Organization for Standardization, ISO 3382-1:2009 Acoustics — Measurement of room acoustic parameters — Part 1: Performance spaces, 2009, (International Standard). ISO 3382-1:2009, https://www.iso.org/standard/40979.html.