KU LEUVEN

STADIUS
Center for Dynamical Systems, Signal Processing and Data Analytics

Citation/Reference	Authors: Kathleen MacWilliam, Thomas Dietzen, Toon van Waterschoot Title: Room Geometry Inference from Room Impulse Responses via First-Order Image-Source Optimization Journal: IEEE Transactions on Audio, Speech, and Language Processing
Archived version	■ Submitted version ☐ Accepted (Author manuscript: the content is identical to the content of the published paper, but without the final typesetting by the publisher) ☐ Published ☐ Published (Open Access)
Acknowledgements	In this manuscript, the following funding is acknowledged: ■ EU H22020 ☐ EU Horizon Europe ■ FWO ■ KULeuven Internal Funds ☐ VLAIO ☐ AI ☐ Other
Author contact	kathleen.macwilliam@esat.kuleuven.be



Room Geometry Inference from Room Impulse Responses via First-Order Image-Source Optimization

Kathleen MacWilliam, Thomas Dietzen, and Toon van Waterschoot

***Abstract*—Estimating room geometry from acoustic measurements is a fundamental problem in spatial audio and room acoustics. Conventional approaches typically rely on detecting and labeling early reflection features, a process that is unreliable and requires strong assumptions for echo sorting. In this work, we propose a reflection feature extraction-free framework for room geometry inference that operates directly on measured room impulse responses (RIRs). Our method leverages an image source model (ISM) parametrized exclusively by first-order image-source positions, from which higher-order reflections can be expressed analytically. We formulate room geometry inference as an optimization problem over these first-order image-source positions, thereby representing room geometry implicitly within a physically motivated RIR model without requiring explicit reflection detection, solving an echo labeling problem (ELP), or reconstructing large sets of higher-order image sources. To reduce sensitivity to amplitude scaling and unknown acoustic parameters, we adopt a cosine-similarity objective that emphasizes temporal structure rather than absolute magnitude. To ensure smooth gradients and stable optimization, we replace bandlimited sinc kernels with temporally localized Gaussian kernels and apply matched smoothing to the measured RIRs. We solve the resulting nonconvex problem using a practical multi-stage procedure that combines multi-start Adam optimization with clustering-based consolidation of stable solutions. In simulations, performance improves with microphone count and spatial aperture and remains robust under moderate sensor noise and reduced boundary reflectivity. On measured RIRs from two datasets, the method recovers a majority of first-order image sources despite using roughly estimated uniform boundary reflection coefficients. Overall, the approach provides a simple, ELP-free pathway to reliable first-order image-source recovery in general convex rooms and offers a solid foundation for further boundary parameter estimation.**

***Index Terms*—Acoustic room impulse responses, room inference, echo labeling problem**

I. Introduction

The ability to recover the positions and orientations of room boundaries is important for a wide range of audio and spatial signal-processing applications, including spatial audio rendering [1], telepresence and teleconferencing [2], virtual and augmented reality [3], and acoustic scene analysis [4]. Accurate geometric knowledge is a key enabler for realistic soundfield synthesis [5], improved dereverberation and echo cancellation [6], [7], robust source localization [8], and systematic design of microphone and loudspeaker placements [9].

Using acoustic measurements to recover this room information is of particular interest because it has a few advantages over vision and LiDAR based approaches. Specifically, microphones do not require line of sight, work reliably in visually obscured environments, and naturally capture acoustic reflections that are primarily shaped by large planar surfaces such as walls. In many practical settings, these acoustically dominant reflectors are more relevant to downstream audio processing tasks than visually detailed geometry, making acoustic inference more efficient and cost-effective [10]. Moreover, estimating orientation directly relative to the microphone array eliminates the additional calibration steps required when using separate measurement devices.

The majority of literature approaches acoustic room inference by assuming that early reflections can first be represented by estimated acoustic parameters such as times of arrival (TOAs) [11]–[27], time differences of arrival (TDOAs) [28], and/or directions of arrival (DOAs) [29]–[31]. We refer to these estimated quantities collectively as reflection features, which are predominantly extracted from room impulse responses (RIRs) encoding the temporal arrival structure of the direct sound and subsequent early reflections. Once reflection feature information has been obtained and correctly associated with its corresponding reflector(s), geometric inversion or room acoustic model–based reasoning can be used to recover reflector positions. Most existing work relies, either explicitly or implicitly, on the image source model (ISM), under which each early reflection corresponds to a virtual so-called image source whose position is determined by the room's boundary planes [32]. Recovering the room geometry thus reduces to identifying the first-order image sources, from which the reflectors can be recovered as planes bisecting the line segments connecting each of their positions with that of the original source.

Methods that rely on extracting specific reflection features face two main challenges. First, they typically require solving a combinatorial echo-labeling problem (ELP) to associate each detected arrival with its corresponding reflector(s) prior to localization. Second, their performance depends strongly on reliable reflection feature detection. Numerous strategies have been proposed to address the ELP, including graph-theoretic formulations [23], greedy assignment schemes [24], [33], Euclidean-distance-matrix (EDM) consistency constraints [13], [23], and convex-hull expansion techniques [15]. Despite these advances, the ELP scales poorly in practical multi-microphone settings, as improving geometric

All authors are with the Department of Electrical Engineering (ESAT), KU Leuven, Leuven, Belgium. Kathleen MacWilliam and Toon van Waterschoot are affiliated with the STADIUS unit, and Thomas Dietzen is with the PSI/EAVISE unit. This research work was carried out at the ESAT Laboratory of KU Leuven, in the frame of KU Leuven internal funds C3/23/056 and C14/21/075, FWO SBO Project S005525N, and FWO Research Project G0A0424N. The research leading to these results has received funding from the Flemish Government (AI Research Program), and from the European Research Council under the European Union's Horizon 2020 research and innovation program / ERC Consolidator Grant: SONORA (no. 773268). This paper reflects only the authors' views and the Union is not liable for any use that may be made of the contained information.

accuracy often requires additional microphones, leading to a rapid combinatorial growth in possible reflection associations and rendering exhaustive labeling impractical. Some approaches reduce labeling ambiguity by incorporating DOA information [29], while others bypass explicit labeling by performing brute-force searches over discretized reflector configurations and matching predicted and observed measurements from uncalibrated RIRs using robust nearest-neighbor cost functions [14]. Several works further assume oracle labeled TOA estimates [34], carefully placed sources [12], [35], or favorable array geometries [36], which limits applicability in realistic and unconstrained environments. Reflection features are commonly extracted from RIRs via peak picking [37], [38], sparse arrival detection [39], matched filtering, or array- and spherical-harmonic-based processing [29], [40]. However, reliable detection becomes challenging when reflections overlap temporally, when microphone arrays are spatially compact, or when the signal-to-noise ratio is low [8], [39]. Inaccuracies at this stage directly impact the ability to solve the ELP.

To reduce reliance on explicit reflection feature extraction, sparse recovery formulations model the RIR as a superposition of virtual or equivalent sources placed on a discretized delay grid. The resulting sparse linear system can be solved with ℓ_1 minimization [28] or compressive sensing techniques [41]. While such methods can improve robustness to noise and overlapping arrivals, grid-based formulations introduce basis mismatch [42] and suffer from rapidly increasing computational and memory requirements in three-dimensional settings.

More recently, several approaches have performed ISM-based inversion for image-source estimation without first extracting acoustic features. Treating the RIR as the output of an ISM forward model, these methods estimate large sets of higher-order image sources by minimizing the error between measured and predicted RIRs [43], [44]. Such approaches directly exploit the full temporal structure of the RIR and naturally handle overlapping reflections. While the method in [44] still solves this in a grid-based framework, [43] addresses the associated limitations by estimating image source locations directly in continuous space without predefined spatial grids. In particular, the authors in [43] reformulate echo localization as a convex optimization problem over Radon measures, regularized with a total variation norm and solved using the Sliding Frank-Wolfe algorithm, enabling high-resolution recovery of large numbers of three-dimensional image sources. Complementary theoretical results on the full reversal of the shoebox ISM establish identifiability for extensive families of image sources [45]. While [44] and [43] successfully estimate large sets of image sources, they both still require a secondary labeling step to actually extract first-order image sources. [44] does this by imposing a compact microphone array and structural constraints derived from the ISM, similarly [45] exploits the structured geometry of image sources by scoring candidate vectors according to orthogonality relationships induced by image-source pairs, but this is limited to box-shaped rooms only and has not been validated on real data.

Lastly, a variety of data-driven and hybrid approaches have been explored [46]. For example, deep neural networks have been employed to directly regress room geometry and surface reflection coefficients from a single RIR without requiring known source/receiver locations, using convolutional neural networks (CNNs) for geometry and MLPs for material properties [47]. Building on RIR structure, other work simultaneously estimates source and receiver positions, room dimensions, and reflection coefficients by analysing the TOAs of direct and reflected components [48]. Extensions to multichannel settings use learned feature extractors to infer 3D room shape from RIRs captured with spherical arrays [49]. In related data-driven reflector localization frameworks, recurrent neural networks (RNNs) process beamforming maps derived from multi-channel RIRs to jointly detect and localize individual reflective surfaces [50], and similar supervised models infer room geometry from RIRs measured with a linear loudspeaker array and single microphone [51]. While these learned approaches often achieve strong empirical performance and can handle complex reverberation patterns, they typically rely on large simulated training sets and may struggle to generalize to arbitrary room shapes, sensor layouts, or acoustical conditions not seen during training.

In this work, we adopt an ISM-based inversion framework and propose a method that directly optimizes a cost function with RIRs parametrized solely in terms of first-order image sources. By optimizing directly over these first-order image sources, our formulation eliminates the need for reflection feature extraction, echo labeling, and higher-order image-source reconstruction. Leveraging the structural information embedded in measured RIRs reduces the problem dimensionality to $3I$ unknowns, where I is the number of reflectors. This simplification facilitates optimization and enables robust recovery of reflector positions in general convex polyhedral rooms, where the ISM provides a reasonable approximation. The key contributions of this work are:

1) We introduce a differentiable RIR model parametrized exclusively by first-order image sources, allowing direct optimization of the error between measured and predicted RIRs without intermediate reflection feature detection or echo labeling.
2) We develop a practical multi-stage optimization and clustering procedure that robustly estimates first-order image source positions, even in the presence of overlapping reflections and moderate noise.
3) We validate the approach extensively through simulations and measured RIRs, demonstrating decent real-world performance and providing an initial assessment of its robustness in practical settings.

The remainder of the paper is structured as follows. Sec. II presents the signal model and problem formulation. Sec. III introduces the first-order image-source parametrized RIR model and formulates the optimization problem, while Sec. IV details the proposed algorithm and the procedure for first-order image source recovery. Experimental validation is provided in Sec. V, with quantitative results and discussion in Sec. VI. Finally, Sec. VII concludes the paper.

II. Signal Model and Problem Statement

Consider a room containing a stationary source located at $\mathbf{p}_0 \in \mathbb{R}^3$, and a set of microphones located at $\mathbf{r}_m \in \mathbb{R}^3$ for $m = 1, \ldots, M$, where m indexes each microphone. An RIR is the time-domain function that characterizes the acoustic propagation between the source and a microphone under the assumption of a linear time-invariant system.

The RIR captures both the direct path from the source to the microphone and the contributions from reflections within the room. A widely adopted model for simulating these reflections is the ISM [32]. In the ISM, each reflection is modeled by a virtual image-source obtained by mirroring the real source across the boundary planes, yielding the set of first-order image sources. Higher-order reflections are then modeled by recursively mirroring these first-order images across additional planes, so that a second-order image is obtained by reflecting a first-order one, a third-order image by reflecting a second-order one, and so on. Although the ISM becomes inaccurate in the late reverberant tail due to scattering and diffusion effects [13], [52], it provides a sufficiently accurate and physically interpretable model for the early reflections that are the focus of this work [53]. Formally, let

$$\mathcal{R} = \{0, 1, \ldots, R-1\} \tag{1}$$

denote the index set of the original source and image sources, and let

$$\mathcal{P} = \{\mathbf{p}_s \in \mathbb{R}^3 : s \in \mathcal{R}\} \tag{2}$$

collect their positions, where $s = 0$ corresponds to the original source. Under the ISM, the continuous-time RIR between a source and microphone m is modeled as a sum of delayed and scaled Dirac impulses corresponding to the direct and reflected propagation paths:

$$h_m(t; \mathcal{P}) = \sum_{s \in \mathcal{R}} a_{s,m}(\mathbf{p}_s)\, \delta\Big(t - \frac{d_{s,m}(\mathbf{p}_s)}{c}\Big), \tag{3}$$

where $c = 343$ m/s is the speed of sound and

$$d_{s,m}(\mathbf{p}_s) = \|\mathbf{r}_m - \mathbf{p}_s\|_2 \tag{4}$$

is the propagation distance from (image) source s to microphone m. The corresponding amplitude $a_{s,m}(\mathbf{p}_s)$, accounts for both free-space attenuation and wall reflections, and is defined here as

$$a_{s,m} = \begin{cases} \dfrac{1}{4\pi d_{s,m}}, & s = 0, \\ \dfrac{\prod_{i \in \mathcal{W}_s} \Gamma_i}{4\pi d_{s,m}}, & s \geq 1. \end{cases} \tag{5}$$

where $\mathcal{W}_s$ denotes the set of walls involved in generating the s-th image source, and $\Gamma_i \in [-1, 1]$ is the scalar, frequency-independent reflection coefficient of wall i. Allowing Γ_i to take negative values enables phase inversion at soft or compliant boundaries, whereas positive values correspond to rigid or partially absorbing surfaces. This formulation can be interpreted as a mild generalization of the classical ISM of [32], in which wall reflections are typically modeled using fixed, non-negative coefficients. More elaborate extensions of the ISM — such as frequency-dependent boundary behaviour implemented via convolutive wall impulse responses, or direction-dependent source and microphone directivity patterns [53] — are not considered here. While such effects have been shown to be important for highly realistic RIR synthesis, the present formulation is sufficient for robustly modelling early reflections in measured RIRs within the scope of the proposed method.

To obtain a discrete-time representation, we impose a bandlimit and sample with period T. Bandlimiting a Dirac delta yields a sinc function [54], sampled at discrete indices. Thus, the discrete-time RIR at microphone m and time index n can be given by

$$h_m(n; \mathcal{P}) = \sum_{s \in \mathcal{R}} a_{s,m}(\mathbf{p}_s)\, \operatorname{sinc}\Big(n - \frac{d_{s,m}(\mathbf{p}_s)}{cT}\Big). \tag{6}$$

The geometry of the room is implicitly encoded in the RIR, since the source and boundary configuration determine the image source positions. Our objective is to invert this relationship: given measured RIRs, we seek to recover the room boundaries. Specifically, we aim to estimate the locations of just the first-order image sources, as if these are known, the boundaries can simply be reconstructed by computing the bisecting planes of the line segments connecting the original source and each first-order image source. A polygonal room with I boundaries is then represented as the intersection of I half-spaces defined by these planes. To isolate the first-order image sources, we define a subset of indices

$$\mathcal{R}^{(1)} \subseteq \mathcal{R}, \tag{7}$$

corresponding to the $I = |\mathcal{R}^{(1)}|$ first-order image sources. The associated first-order positions are collected in

$$\mathcal{Q} = \{\mathbf{q}_i \in \mathbb{R}^3 : i \in \mathcal{R}^{(1)}\}, \tag{8}$$

which forms a subset of the full set of (image) source positions $\mathcal{P}$, i.e., $\mathcal{Q} \subseteq \mathcal{P}$. Accurate estimation of first-order image sources simplifies subsequent tasks such as echo labeling, attenuation estimation, and reflection coefficient computation. In essence, recovering these positions provides a foundation for comprehensive room geometry and acoustic parameter estimation. Throughout this work we make the following assumptions:

1) The source and microphones are omnidirectional.
2) The room is an empty convex polyhedral enclosure with specular boundary reflections, so the ISM accurately describes the early reflection structure.
3) The relative positions of the source and microphones, and the number of room boundary planes, are known.

III. Proposed Model

We aim to estimate the set of first-order image-source positions $\mathcal{Q} = \{\mathbf{q}_i \in \mathbb{R}^3 : i \in \mathcal{R}^{(1)}\}$ from discrete-time RIRs measured at microphone positions $\mathbf{r}_m$, $m = 1, \ldots, M$. The key observation underlying our approach is that all higher-order image source positions can be expressed as functions of first-order image source positions. This allows us to parametrize the entire early-reflection structure in terms of $\mathcal{Q}$. Rather than extracting and labeling reflection TOAs (which is often unreliable) we fit a forward ISM-based RIR model that embeds the dependence on $\mathcal{Q}$ directly.

A. First-order Image-source Parametrized Forward Model

Without loss of generality, we apply a global coordinate shift such that the original source is located at the origin, i.e., $\mathbf{p}_0 = \mathbf{0}$. Let the boundary plane $i \in \mathcal{R}^{(1)}$ be described using unit outward normal vector $\mathbf{n}_i$ and an arbitrary point on the plane $\mathbf{w}_i$. The corresponding first-order image source can then be obtained by the Householder reflection [55] across boundary i and is given by

$$\mathbf{q}_i = \mathbf{p}_0 - 2\,\mathbf{n}_i^\top(\mathbf{p}_0 - \mathbf{w}_i)\,\mathbf{n}_i = 2\,(\mathbf{n}_i^\top \mathbf{w}_i)\,\mathbf{n}_i. \tag{9}$$

Thus, $\mathbf{q}_i$ is collinear with $\mathbf{n}_i$, and the orientation of $\mathbf{n}_i$ is chosen such that $\mathbf{n}_i = \mathbf{q}_i/\|\mathbf{q}_i\|_2$. We now consider the second-order image source associated with the ordered boundary pair (i,j), where $i,j \in \mathcal{R}^{(1)}$ and $i \neq j$. This ordered pair corresponds to a reflection first at boundary i and subsequently at boundary j. Let $\mathbf{n}_j$ and $\mathbf{w}_j$ denote the unit outward normal vector and an arbitrary point on boundary j, respectively. Reflecting the first-order image source $\mathbf{q}_i$ across boundary j yields

$$\mathbf{q}_{ij} = \mathbf{q}_i - 2\,\mathbf{n}_j^\top\big(\mathbf{q}_i - \mathbf{w}_j\big)\,\mathbf{n}_j \tag{10}$$

$$= \mathbf{q}_i + 2\,(\mathbf{n}_j^\top \mathbf{w}_j)\,\mathbf{n}_j - 2\,(\mathbf{n}_j^\top \mathbf{q}_i)\,\mathbf{n}_j. \tag{11}$$

Since the first-order image source corresponding to boundary j satisfies $\mathbf{q}_j = 2\,(\mathbf{n}_j^\top \mathbf{w}_j)\,\mathbf{n}_j$, and $\mathbf{n}_j = \frac{\mathbf{q}_j}{\|\mathbf{q}_j\|_2}$, the second-order image source associated with the ordered pair (i,j) can be expressed solely in terms of the two first-order image sources as

$$\mathbf{q}_{ij} = \mathbf{q}_i + \mathbf{q}_j - 2\,\frac{\mathbf{q}_i^\top \mathbf{q}_j}{\|\mathbf{q}_j\|_2^2}\,\mathbf{q}_j. \tag{12}$$

In general, the ordered reflection sequences (i,j) and (j,i) with $i \neq j$ correspond to distinct second-order image sources, as they represent different reflection paths. However, in special geometries, such as rectangular rooms with mutually orthogonal boundaries, these image sources may coincide geometrically. By iteration, third-order image sources are defined for ordered triples (i,j,k) as

$$\mathbf{q}_{ijk} = \mathbf{q}_{ij} + \mathbf{q}_k - 2\frac{\mathbf{q}_{ij}^\top \mathbf{q}_k}{\|\mathbf{q}_k\|_2^2}\,\mathbf{q}_k. \tag{13}$$

Importantly, third-order sequences satisfy $i \neq j$ and $j \neq k$, while repetition such as $i = k$ is allowed, corresponding to physically valid multi-bounce paths (e.g., reflection on boundary i, then j, then again i). In the final model used in this work, however, we adopt a simplified indexing convention and restrict attention to sequences with distinct indices, denoted compactly by $i \neq j \neq k$, thereby excluding repeated reflections on the same boundary. In addition, the above construction treats room boundaries as infinitely extended planes; in practice, a reflection contributes physically only if the corresponding reflection points lie within the finite extent of the wall surfaces. Neither repeated-boundary paths nor finite-surface visibility constraints are explicitly enforced in the proposed forward model. Instead, their effects are handled implicitly within the subsequent similarity-based optimization procedure.

Furthermore, we restrict attention to at most third-order reflections, which suffice to capture the early-reflection structure of interest. We now define the propagation distances

$$d_{i,m}^{(1)}(\mathbf{q}_i) = \|\mathbf{r}_m - \mathbf{q}_i\|_2, \tag{14}$$

$$d_{ij,m}^{(2)}(\mathbf{q}_{ij}) = \|\mathbf{r}_m - \mathbf{q}_{ij}\|_2, \tag{15}$$

$$d_{ijk,m}^{(3)}(\mathbf{q}_{ijk}) = \|\mathbf{r}_m - \mathbf{q}_{ijk}\|_2, \tag{16}$$

where (i,j,k), with $i,j,k \in \mathcal{R}^{(1)}$, are ordered indices as defined above. These distances are then used to directly embed the dependence on the first-order image-source positions into the ISM representation of the RIR in Eq. (6). The resulting modeled discrete-time RIR at sample index n and sampling period T, truncated to third-order reflections, is then given by

$$\begin{aligned} h_m(n;\mathcal{Q}) = \frac{1}{4\pi}\Bigg[& \underbrace{\frac{\operatorname{sinc}\Big(n - \frac{\|\mathbf{r}_m\|_2}{cT}\Big)}{\|\mathbf{r}_m\|_2}}_{\text{Direct path}} \\ & + \sum_{i\in\mathcal{R}^{(1)}} \underbrace{\frac{\Gamma_i\,\operatorname{sinc}\Big(n - \frac{d_{i,m}^{(1)}(\mathbf{q}_i)}{cT}\Big)}{d_{i,m}^{(1)}(\mathbf{q}_i)}}_{\text{1st-order reflections}} \\ & + \sum_{\substack{i,j\in\mathcal{R}^{(1)}\\ i\neq j}} \underbrace{\frac{\Gamma_i\Gamma_j\,\operatorname{sinc}\Big(n - \frac{d_{ij,m}^{(2)}(\mathbf{q}_{ij})}{cT}\Big)}{d_{ij,m}^{(2)}(\mathbf{q}_{ij})}}_{\text{2nd-order reflections}} \\ & + \sum_{\substack{i,j,k\in\mathcal{R}^{(1)}\\ i\neq j\neq k}} \underbrace{\frac{\Gamma_i\Gamma_j\Gamma_k\,\operatorname{sinc}\Big(n - \frac{d_{ijk,m}^{(3)}(\mathbf{q}_{ijk})}{cT}\Big)}{d_{ijk,m}^{(3)}(\mathbf{q}_{ijk})}}_{\text{3rd-order reflections}} \Bigg]. \end{aligned} \tag{17}$$

In practice, wall-specific coefficients Γ_i are rarely known a priori. However, for geometry recovery, approximate amplitudes suffice since the geometry is primarily encoded in the relative timing of early reflections. We therefore adopt a single broadband coefficient, $\Gamma_i = \Gamma$ for all i. We estimate Γ from the measured RIR via a direct-to-first-order ratio with spherical spreading compensation:

$$\hat{\Gamma} \approx \frac{\hat{a}_{i,m}^{(1)}}{\hat{a}_{0,m}} \cdot \frac{\hat{d}_{i,m}^{(1)}}{\hat{d}_{0,m}}, \tag{18}$$

where $\hat{a}_{i,m}^{(1)}$ and $\hat{d}_{i,m}^{(1)}$ are the measured amplitude and inferred distance of a first-order reflection, and $\hat{a}_{0,m}$, $\hat{d}_{0,m}$ are the direct-path counterparts[1]. Alternatively, when coarse room data are available, one can estimate Γ from global decay metrics [2] (e.g., reverberation time via Schroeder integration and Sabine/Eyring theory [56]).

[1] In a convex polygonal room, the first two arrivals after the direct path must be first-order reflections; higher orders are strictly longer. Thus, at least two first-order amplitudes in each RIR are available without explicit higher-order labeling and we can take the average of them across different microphone positions to find an estimate of Γ.

[2] Either estimate is sufficiently accurate for this application; surface-specific coefficients can be refined after recovering $\mathcal{Q}$.

B. Objective Formulation

We next define notation for the vectors used in the objective function. Let $\hat{h}_m(n)$ denote the measured RIR at microphone m, and collect the first N samples into

$$\hat{\mathbf{h}}_m = [\,\hat{h}_m(0), \ldots, \hat{h}_m(N-1)\,]^\top \in \mathbb{R}^N. \tag{19}$$

Using the modeled RIR $h_m(n;\mathcal{Q})$ in Eq. (17), we similarly define

$$\mathbf{h}_m(\mathcal{Q}) = [\,h_m(0;\mathcal{Q}), \ldots, h_m(N-1;\mathcal{Q})\,]^\top. \tag{20}$$

A conventional least-squares fit $\min_{\mathcal{Q}} \sum_m \|\hat{\mathbf{h}}_m - \mathbf{h}_m(\mathcal{Q})\|_2^2$ is sensitive to absolute amplitude errors (e.g., imperfect Γ), which can dominate the cost and obscure timing structure. We therefore compare normalized RIRs using cosine similarity:

$$\min_{\mathcal{Q}} \sum_{m=1}^{M} \left(1 - \frac{\hat{\mathbf{h}}_m^\top \mathbf{h}_m(\mathcal{Q})}{\|\hat{\mathbf{h}}_m\|_2 \, \|\mathbf{h}_m(\mathcal{Q})\|_2}\right). \tag{21}$$

This objective suppresses amplitude mismatch while preserving inter-reflection timing and structure.

1) Gaussian Kernel Replacement: The objective in (21) is highly nonconvex in $\mathcal{Q}$, owing to the nonlinear dependence of path lengths on image-source positions, the combinatorial coupling across reflection orders, and the presence of sinc interpolation kernels. While the sinc kernel in the discrete-time ISM model in Eq. (17) arises naturally from bandlimiting and provides an exact representation of fractionally delayed impulses, it poses significant challenges for optimization.

Specifically, the sinc function has infinite support, oscillatory sidelobes, and slow decay, so even small perturbations in an image-source distance —and hence in the associated fractional delay— affect a large number of time samples. Moreover, its derivatives are themselves oscillatory, leading to poorly conditioned and highly variable gradients with respect to image-source positions. When aggregated across microphones and time, these effects accumulate, producing unstable update directions and a strongly nonconvex optimization landscape with numerous undesirable local minima.

In numerical implementations, the sinc kernel must also be truncated or windowed. Truncation limits the temporal support but does not remove the oscillations, while windowing reduces the oscillations but introduces spectral distortion. Obtaining smooth and stable gradients therefore requires careful tuning and often large windows. To avoid these issues, we replace the sinc kernel with a Gaussian kernel, which is smooth, non-oscillatory, and rapidly decaying. Its width is controlled by a single parameter σ, which directly determines the temporal spread. The modified RIR model vector, denoted $\mathbf{h}_m(\mathcal{Q};\sigma)$, is obtained by substituting each sinc term in (17) with

$$\operatorname{sinc}\!\left(n - \tfrac{d}{cT}\right) \;\longrightarrow\; \exp\!\left(-\frac{(n - \frac{d}{cT})^2}{2\sigma^2}\right), \tag{22}$$

where $\sigma > 0$ sets the kernel width. This substitution preserves reflection timing while producing localized kernels with smoother derivatives, resulting in a more stable dependence on $\mathcal{Q}$ that is better suited to gradient-based optimization.

To ensure consistency between the model and data, the measured RIRs $\hat{\mathbf{h}}_m$ are smoothed using a discrete Gaussian kernel whose width matches the Gaussian kernels employed in the forward model. For each measured RIR, we construct a symmetric kernel with support $[-3\sigma, 3\sigma]$ and normalize its coefficients to unit sum:

$$g(k) = \frac{\exp(-k^2/(2\sigma^2))}{\sum_{\ell=-3\sigma}^{3\sigma} \exp(-\ell^2/(2\sigma^2))}. \tag{23}$$

The smoothed RIR is then obtained by convolution,

$$\hat{h}_m^{(\sigma)}(n) = \sum_{k=-3\sigma}^{3\sigma} \hat{h}_m(n-k)\, g(k). \tag{24}$$

The effect of the Gaussian kernel width σ on the smoothing of a sinc-simulated RIR at a sampling frequency of 16 kHz is illustrated in Fig. 1. For visualization purposes, each RIR is rescaled to match the maximum amplitude of the original RIR (i.e., the direct-path arrival), compensating for the reduction in peak magnitude caused by smoothing. It can be observed that increasing σ progressively broadens individual reflections and attenuates high-frequency oscillations, while preserving the overall temporal structure of the RIR[3]. With the Gaussian model and the smoothed measurements, the cosine similarity objective function is finally defined as

$$J_{\cos}(\mathcal{Q}) = \sum_{m=1}^{M} \left(1 - \frac{\hat{\mathbf{h}}_m^{(\sigma)\top} \mathbf{h}_m(\mathcal{Q};\sigma)}{\|\hat{\mathbf{h}}_m^{(\sigma)}\|_2 \, \|\mathbf{h}_m(\mathcal{Q};\sigma)\|_2}\right), \tag{25}$$

where $\mathbf{h}_m(\mathcal{Q};\sigma)$ is the Gaussian-kernel ISM RIR vector and $\hat{\mathbf{h}}_m^{(\sigma)}$ is the smoothed measured RIR vector. To illustrate the effect of this replacement on the optimization landscape, Fig. 2 shows representative two-dimensional slices of the objective function evaluated over the (x, y) coordinates of a single first-order image source, with all remaining image-source locations held fixed at their ground-truth values. The contour lines therefore represent level sets of equal cost, corresponding to local cross-sections of the full high-dimensional objective around the true solution. Compared to the sinc-based formulation, the Gaussian model yields a smooth surface with reduced oscillatory structure and a well-defined basin of attraction

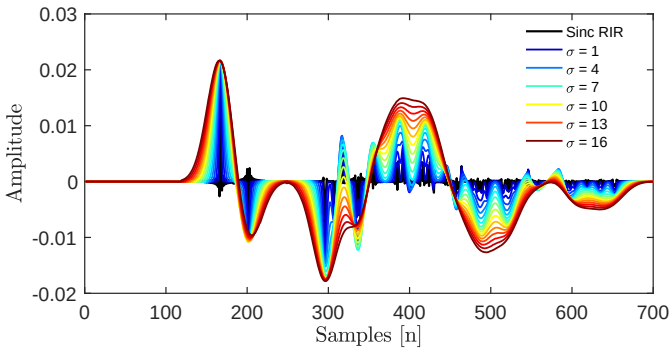

Fig. 1: Effect of Gaussian smoothing on a sinc-simulated RIR with a sampling frequency of 16 kHz for increasing values of σ. The original RIR is shown in black, with colored curves corresponding to progressively larger Gaussian kernel widths.

[3]This behavior follows from the fact that convolution with a Gaussian in the time domain corresponds to multiplication by a Gaussian in the frequency domain, whose bandwidth is inversely proportional to σ; larger values of σ therefore increasingly suppress high-frequency content, whereas very small $\sigma \ll 1$ retain fine temporal detail but result in minimal low-pass filtering.

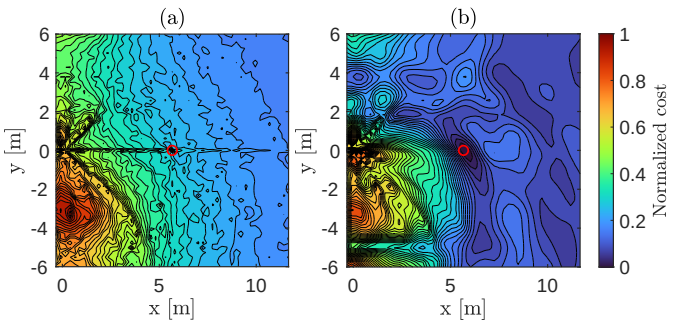

Fig. 2: Representative (x, y) slices of the objective function for a single unknown first-order image source, with all other first-order image-source positions held fixed. A sampling frequency of 16 kHz is used. The contour lines indicate regions of equal cost in the objective function. The red circle denotes the true first-order image-source location. (a) Sinc-based model. (b) Gaussian-kernel model with $\sigma = 16$.

around the true image-source location. In practice, σ must balance temporal resolution against optimization stability: small σ values yield sharp but more oscillatory objective functions with poorly conditioned gradients, while larger σ values produce smoother objective landscapes and improved convergence at the cost of temporal precision. At a sampling rate of 16 kHz, we found $\sigma = 16$ samples to provide a good compromise between these effects, yielding reliable convergence while maintaining sufficient localization accuracy.

2) Repulsive Gaussian Penalty: Finally, as an additional helper we optionally include a smooth repulsive regularization term defined over all distinct pairs of first-order image-source positions:

$$J_{\text{rep}}(\mathcal{Q}) = \lambda \sum_{\substack{i,j \in \mathcal{R}^{(1)} \\ i \neq j}} \exp\left(-\frac{\|\mathbf{q}_i - \mathbf{q}_j\|_2^2}{2\sigma_{\text{rep}}^2}\right), \tag{26}$$

where $\lambda > 0$ controls the overall strength of the penalty and $\sigma_{\text{rep}} > 0$ defines a soft minimum separation distance [4]. The Gaussian form ensures that the regularizer is fully differentiable and interacts smoothly with the main objective.

3) Final Objective Formulation: Combining the terms above, our estimation of $\mathcal{Q}$ is posed as

$$\boxed{\min_{\mathcal{Q}} \; J(\mathcal{Q}) \;=\; \min_{\mathcal{Q}} \; (J_{\cos}(\mathcal{Q}) \;+\; J_{\text{rep}}(\mathcal{Q}))} \tag{27}$$

where $J_{\cos}(\mathcal{Q})$ and $J_{\text{rep}}(\mathcal{Q})$ are defined in (25) and (26), respectively. This section has specified the forward model and objective; the next chapter details the algorithm used to solve (27).

IV. Proposed Multi-Stage Adam–Clustering Refinement Algorithm

For numerical optimization, we parametrize the image-source positions by stacking the elements of $\mathcal{Q}$ into a single vector

$$\mathbf{x} = \left[\mathbf{q}_i^\top\right]_{i \in \mathcal{R}^{(1)}}^\top \in \mathbb{R}^{3I}. \tag{28}$$

[4]Physically, first-order image sources only become close when the actual source lies very near room boundaries or corners. Accordingly, σ_{rep} is not intended to encode a precise geometric prior, but rather serves as a coarse regularization scale. In our experiments, we use a conservative value of $\sigma_{\text{rep}} \approx 1\,\text{m}$, which is appropriate for typical room configurations.

With slight abuse of notation, the cost function $J(\mathcal{Q})$ defined in (27) can equivalently be written as

$$J(\mathbf{x}) := J(\mathcal{Q}). \tag{29}$$

Optimization is therefore performed with respect to $\mathbf{x} \in \mathbb{R}^{3I}$. Due to the non-convex nature of (27), a single local optimization is unlikely to recover a globally meaningful first-order image-source configuration. We therefore propose a multi-stage refinement strategy that combines repeated multi-start Adam optimization with clustering-based consolidation of stable solutions. Each refinement stage is formally identical and differs only in its initialization, which is informed by the consolidated solutions obtained in the preceding stage.

A. Stage-wise Multi-start Optimization

At refinement stage $l = 1, \ldots, L$, we generate H independent candidate initializations

$$\mathbf{x}_0^{(l,h)} = \left[\mathbf{q}_i^{(l,h)\top}\right]_{i \in \mathcal{R}^{(1)}}^\top \in \mathbb{R}^{3I}, \qquad h = 1, \ldots, H, \tag{30}$$

where $\mathbf{q}_i^{(l,h)}$ denotes the current estimate of the i-th first-order image-source position for candidate h at stage l. Each candidate is then optimized independently using Adam [57] for F fixed iterations. For $f = 1, \ldots, F$, the updates are

$$\mathbf{g}_f^{(l,h)} = \nabla_{\mathbf{x}} J(\mathbf{x}_{f-1}^{(l,h)}), \tag{31}$$

$$\mathbf{m}_f^{(l,h)} = \beta_1 \mathbf{m}_{f-1}^{(l,h)} + (1-\beta_1)\mathbf{g}_f^{(l,h)}, \tag{32}$$

$$\mathbf{v}_f^{(l,h)} = \beta_2 \mathbf{v}_{f-1}^{(l,h)} + (1-\beta_2)\big(\mathbf{g}_f^{(l,h)}\big)^2, \tag{33}$$

$$\widehat{\mathbf{m}}_f^{(l,h)} = \frac{\mathbf{m}_f^{(l,h)}}{1-\beta_1^f}, \qquad \widehat{\mathbf{v}}_f^{(l,h)} = \frac{\mathbf{v}_f^{(l,h)}}{1-\beta_2^f}, \tag{34}$$

$$\mathbf{x}_f^{(l,h)} = \mathbf{x}_{f-1}^{(l,h)} - \alpha \frac{\widehat{\mathbf{m}}_f^{(l,h)}}{\sqrt{\widehat{\mathbf{v}}_f^{(l,h)}} + \varepsilon}. \tag{35}$$

Here, α is the learning rate, $\beta_1, \beta_2 \in (0,1)$ are exponential decay factors, and $\varepsilon > 0$ ensures numerical stability. The $(.)^2$ and $\sqrt{.}$ operations are applied elementwise. The moment vectors are initialized as $\mathbf{m}_0^{(l,h)} = \mathbf{0}$ and $\mathbf{v}_0^{(l,h)} = \mathbf{0}$, while the initialization of $\mathbf{x}_0^{(l,h)}$ is discussed in Sec. IV-B. Each run produces a locally optimized configuration $\mathbf{x}_F^{(l,h)}$.

B. Initialization

Without loss of generality, let $\mathbf{r}_1 \in \mathbb{R}^3$ denote a designated reference microphone. From the RIR measured at $\mathbf{r}_1$, we extract a set of coarse TOAs corresponding to the I most prominent peaks, convert them to propagation distances, and store them as $\{D_i\}_{i=1}^I$. These TOAs are obtained using a simple and intentionally approximate preprocessing stage (matched filtering followed by a generic peak-picking routine such as in [12]) and are not required to correspond to true first-order reflections. Based on these distances, we define a set of spheres

$$\mathbb{S}_i := \left\{\mathbf{q} \in \mathbb{R}^3 : \|\mathbf{q} - \mathbf{r}_1\|_2 = D_i\right\}, \tag{36}$$

which serve as soft geometric priors for initialization. At stage $l = 1$, each candidate $\mathbf{x}_0^{(1,h)}$ is formed by randomly sampling

$$\mathbf{q}_i^{(1,h)} \sim \mathrm{Unif}(\mathbb{S}_i), \quad i = 1, \ldots, I, \tag{37}$$

where the notation $\sim$ denotes a random draw from the specified distribution, and $\mathrm{Unif}(\mathbb{S}_i)$ denotes the uniform distribution over the continuous surface of the sphere $\mathbb{S}_i$, followed by stacking over i. A minimum pairwise separation constraint is enforced between each pair of candidates to promote spatial diversity. Importantly, this initialization is not essential for the correctness of the proposed algorithm. The refinement procedure can also be initialized using uninformed random sampling in the search volume, in which case it remains convergent but typically exhibits much slower convergence and reduced recall. The role of the TOA-based spheres is therefore to improve conditioning and exploration efficiency by biasing initial samples towards plausible regions of space, rather than to perform accurate reflection feature extraction. Consistent with this interpretation, the extracted peaks need not be reliable or uniquely associated with first-order reflections; they simply provide a coarse spatial spread of candidate distances that accelerates convergence in early stages.

C. Clustering-based Consolidation

After stage l, optimized candidates are pooled as

$$\mathcal{X}_{\mathrm{pool}} = \bigcup_{l'=1}^{l} \{\mathbf{x}_F^{(l',h)}\}_{h=1}^{H}. \tag{38}$$

The κ globally lowest-cost solutions in $\mathcal{X}_{\mathrm{pool}}$ are then selected according to $J_{\cos}(\mathcal{Q})$ and only these are used for the subsequent clustering step, allowing strong solutions from earlier stages to contribute alongside later refinements. All individual 3D coordinate vectors from these retained candidates are aggregated into a point cloud and clustered (e.g., using DBSCAN [58]) with parameters:

1) $\varepsilon_{\mathrm{cluster}}$: maximum spatial distance for cluster membership,
2) $C_{\min}$: minimum number of points for a stable cluster.

Clusters containing at least $C_{\min}$ points are declared stable. Each stable cluster is assigned to the nearest sphere $\mathbb{S}_i$, representing its corresponding first-order image source. In case of multiple clusters assigned to the same sphere, the largest cluster is chosen. For the next stage $l + 1$, each first-order image source associated with a stable cluster is reinitialized at the corresponding cluster centroid across all H candidate initializations, while remaining sources are independently resampled from their respective spheres $\mathbb{S}_i$.

D. Stopping Criteria and Final Estimate

The procedure iterates until one of the following conditions is met:

1) the number of stable clusters reaches I,
2) the number of stable clusters stagnates over η_{stag} consecutive stages, or
3) the maximum stage count L is reached.

Algorithm 1 Multi-Stage Adam–Clustering Refinement

Inputs: J, spheres $\{\mathbb{S}_i\}$, $H, F, \kappa, \varepsilon_{\mathrm{cluster}}, C_{\min}, L, \eta_{\mathrm{stag}}$
Outputs: $\widehat{\mathbf{x}}$

$\mathcal{X}_{\mathrm{pool}} \leftarrow \emptyset$, stagn $= 0$, prev_count $= -1$
for $l = 1$ to L **do**
  for $h = 1$ to H **do**
    if $l = 1$ **then**
      Sample $\mathbf{q}_i \sim \mathrm{Unif}(\mathbb{S}_i)$; enforce minimum separation
    else
      $\mathbf{q}_i \leftarrow$ (cluster centroid if stable, else sample on $\mathbb{S}_i$), $\forall i$
    end if
    $\mathbf{x}_F^{(l,h)} \leftarrow \mathrm{Adam}(J, \{\mathbf{q}_i\}, F)$
  end for
  $\mathcal{X}_{\mathrm{pool}} \leftarrow \mathcal{X}_{\mathrm{pool}} \cup \{\mathbf{x}_F^{(l,h)}\}$
  Retain κ lowest-cost elements
  Form 3D point cloud; run DBSCAN$(\varepsilon_{\mathrm{cl}}, C_{\min})$
  Extract stable clusters and assign one per sphere (keep largest if multiple)
  if $|\text{stable}| = I$ **then break**
  end if
  if $|\text{stable}| =$ prev_count **then**
    stagn $\leftarrow$ stagn $+ 1$
  else
    stagn $\leftarrow 0$
  end if
  prev_count $\leftarrow |\text{stable}|$
  if stagn $\geq \eta_{\mathrm{stag}}$ **then break**
  end if
end for
if $|\text{stable}| = I$ **then**
  $\widehat{\mathbf{x}} \leftarrow \mathrm{Adam}(J, \text{centroids}, F)$
else if $|\text{stable}| > 0$ **then**
  $\widehat{\mathbf{x}} \leftarrow$ stacked centroids
else
  $\widehat{\mathbf{x}} \leftarrow$ lowest-cost element in $\mathcal{X}_{\mathrm{pool}}$
end if
return $\widehat{\mathbf{x}}$

If all I stable clusters are obtained, their centroids initialize a final joint Adam optimization over all coordinates. If only a subset of $I' \leq I$ stable clusters is found, their centroids are returned as the detected first-order image sources. If no stable clusters exist, the lowest-cost candidate in $\mathcal{X}_{\mathrm{pool}}$ is returned. The final estimate is

$$\widehat{\mathbf{x}} \in \mathbb{R}^{3I'}, \qquad I' \leq I, \tag{39}$$

When $I' < I$, the estimate represents a partial reconstruction of the room. Empirically, physically correct image sources form persistent spatial clusters across stages due to repeated reinitialization at consistent locations across candidate runs, whereas spurious local minima fail to stabilize. Pseudocode summarizing the proposed algorithm is presented in Algorithm 1.

V. Experimental Validation

This section outlines the experiments used to validate the proposed algorithm. Both simulated and real RIRs are considered. Simulations enable controlled evaluation and proof of concept, whereas real-data experiments assess robustness in practical settings. Quantitative results are presented in Section VI.

A. Performance Metrics

Estimation performance is evaluated along two complementary dimensions: geometric localization accuracy and detection performance (recall). Recollect that the coordinate frame is translated such that the physical source lies at the origin. For any pair of ground truth and estimated first-order image sources $(\mathbf{q}_i, \widehat{\mathbf{q}}_i)$ we define

$$\epsilon_d = \|\widehat{\mathbf{q}}_i - \mathbf{q}_i\|_2, \qquad \epsilon_r = \big|\|\widehat{\mathbf{q}}_i\|_2 - \|\mathbf{q}_i\|_2\big|,$$
$$\epsilon_\theta = \cos^{-1}\left(\frac{\widehat{\mathbf{q}}_i^\top \mathbf{q}_i}{\|\widehat{\mathbf{q}}_i\|_2 \, \|\mathbf{q}_i\|_2}\right), \tag{40}$$

corresponding to Euclidean, radial, and angular errors. Recorded geometric errors are averaged across correctly detected sources within each trial and then across trials. An estimate is counted as a correct detection if it can be uniquely assigned to a ground-truth image source such that $\epsilon_d < 0.85\,\mathrm{m}$ and $\epsilon_\theta < 10^\circ$. These same (deliberately lenient) tolerances are used for simulated and real experiments to enable direct comparison. Two factors motivate this choice. First, the Gaussian smoothing applied to the RIRs broadens early-reflection peaks, increasing the uncertainty in their inferred arrival times. Second, by the reverse triangle inequality,

$$\big|\|\mathbf{r}_m - \widehat{\mathbf{q}}_i\|_2 - \|\mathbf{r}_m - \mathbf{q}_i\|_2\big| \le \|\widehat{\mathbf{q}}_i - \mathbf{q}_i\|_2, \tag{41}$$

the Euclidean localization error provides a worst-case bound on the corresponding timing deviation at any microphone position $\mathbf{r}_m$:

$$\Delta t \le \frac{\|\widehat{\mathbf{q}}_i - \mathbf{q}_i\|_2}{c}. \tag{42}$$

Thus, errors near the tolerance of $0.85\,\mathrm{m}$ correspond to a maximum timing deviation of approximately $2.48\,\mathrm{ms}$, or about 40 samples at $16\,\mathrm{kHz}$, which remains practically meaningful for many room-geometry applications and early-reflection analysis. Finally, let $\mathcal{Q}_{\text{correct}} \subseteq \mathcal{Q}$ denote the subset of image sources that are successfully matched. The recall is then defined as

$$\text{Recall} = \frac{|\mathcal{Q}_{\text{correct}}|}{I}. \tag{43}$$

B. Algorithmic Parameters

All algorithmic parameters that are fixed across both simulations and real experiments are summarized in Table I. The parameters of the repulsive regularization term in Eq. (26) are set to $\lambda = 1$ and $\sigma_{\text{rep}} = 1\,\mathrm{m}$. The I largest peaks in the RIR from a designated reference microphone are extracted after applying a matched filter and peak picking [12]. These peaks are then used to initialize the spherical search regions described in Eq. (36). Under ideal (noise-free) simulated RIR conditions, the algorithm typically converges within $L \le 3$ stages. For use in the RIR model, a single effective reflection coefficient Γ is estimated from each measured or simulated RIR using the direct-path component and the first two reflection peaks (see footnote 1). In the simulation study, negative reflection coefficients are employed. This choice is motivated by the commonly used acoustic boundary model in which rigid boundaries produce reflections with a phase inversion. In contrast, the reflection coefficients estimated from the measured RIRs in the real-data experiments are predominantly positive. Evaluating both cases therefore demonstrates the robustness of the proposed algorithm under different reflection-coefficient conditions.

TABLE I: Algorithmic parameters used in experiments.

Param.	Value	Description
Adam	$\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$, $F = 50$	Optimizer hyperparameters and iterations per run.
Clustering	$\varepsilon_{\text{cluster}} = 0.2\,\mathrm{m}$, $C_{\min} = 3$, $\kappa = 50$, $\eta_{\text{stag}} = 5$	Radius, min. cluster size, retained configs., and stagnation threshold.
Init./Model	$H = 200$, I, Γ	No. of initial candidates, reflectors, and reflection coefficient.

Two sampling frequencies are also considered. Simulations are performed at $f_s = 16\,\mathrm{kHz}$, which is typical for speech processing applications and reduces computational cost during extensive simulation-based validation. The real recordings, however, are processed at their native sampling rate of $f_s = 48\,\mathrm{kHz}$ in order to preserve the full temporal resolution of the measured RIRs. Finally, the gradient of the general cosine-similarity cost function in Eq. (27) with respect to the first-order image-source positions is derived analytically. For convenience, a concise summary of the derivation, along with a MATLAB implementation of the resulting gradient, is provided in the supplementary material. Portions of the code associated with this work will also be made publicly available online.[5]

C. Simulated Experiments

All simulations are performed in a rectangular room of dimensions $5.78 \times 6.7 \times 4\,\mathrm{m}$. A single source is placed at a fixed off-center location at least $1\,\mathrm{m}$ from each wall. Microphone positions are randomly generated subject to experiment-specific constraints. RIRs are synthesized using the ISM [32] with a sampling rate of $f_s = 16\,\mathrm{kHz}$. Although a shoebox geometry is used for convenience, the proposed model does not rely on this assumption. For computational efficiency, simulated RIRs are limited to second-order reflections, with the model in Eq. (17) modified accordingly. This restriction is adopted solely to reduce runtime in large-scale simulations and does not indicate a limitation of the proposed method in handling third-order reflections. While third-order reflections are included in the real-data experiments, their omission in simulation has negligible impact on the quantitative behavior under evaluation.

*1) Effect of Microphone Count and Gaussian Kernel Width (**Experiment 1**):* This experiment evaluates performance as a function of microphone count, using $M \in \{6, 8, 10, 12, 14, 16\}$. For each value of M, we generate 50 random microphone configurations, with all microphones placed at least $0.5\,\mathrm{m}$ from any surface. Each configuration is tested for three Gaussian kernel widths, $\sigma \in \{4, 8, 16\}$. Simulation wall reflection coefficients are drawn independently from $\Gamma_i \sim -\mathcal{U}(0.8, 0.9)$, where $\mathcal{U}(a, b)$ denotes the continuous uniform distribution on the interval $[a, b]$, representing predominantly hard, planar surfaces. The fitted model assumes

[5] https://github.com/kmacwill/Acoustic-Room-Inference

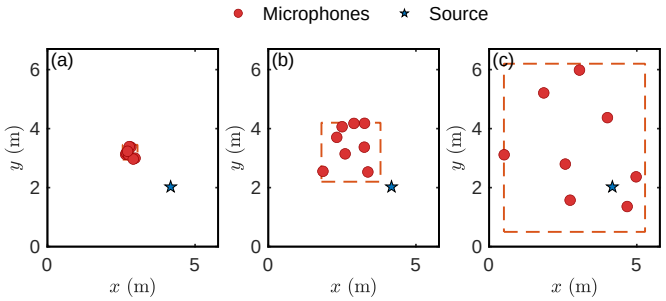

Fig. 3: Example microphone configurations in the xy-plane for three aperture settings: (a) compact $0.5\,\mathrm{m}^3$ region, (b) larger $2\,\mathrm{m}^3$ region, and (c) microphones distributed throughout the full room.

a uniform reflection coefficient of $\Gamma = -0.85$. This experiment establishes a baseline under ideal, noise-free conditions with well-distributed microphones.

*2) Effect of Microphone Aperture (**Experiment 2**):* This experiment examines how performance varies as the arrangement of the microphones varies from compact to widely distributed. We fix the number of microphones at $M = 8$, a relatively small count chosen to reflect realistic deployment scenarios, and vary their spatial aperture. The microphones are initially placed uniformly at random within a compact $0.5\,\mathrm{m}^3$ cube. We then consider progressively larger cubic regions of volume $1\,\mathrm{m}^3$ and $2\,\mathrm{m}^3$, and finally compare these results to the full-room placements from ***Experiment 1***. Fig 3 illustrates example microphone placements in the xy-plane. Representative smoothed RIRs for the corresponding region sizes with $\sigma = 16$ are shown in Fig. 4. For each aperture size, we generate 50 random configurations using the same reflection-coefficient assumptions as in ***Experiment 1***. In addition, we evaluate multiple Gaussian kernel widths $\sigma \in \{4, 8, 16\}$ to observe whether smaller σ values compensate for the reduced spatial diversity present in tightly clustered microphone configurations (see Fig. 4). This allows us to examine the trade-off between microphone aperture and kernel resolution.

*3) Noise and Wall Absorption (**Experiment 3**) :* This experiment evaluates robustness to reduced wall reflectivity and measurement noise. We fix $M = 16$ and generate 50 microphone configurations using the same placement constraints as in ***Experiment 1***. To capture different levels of boundary damping, we consider three simulation reflection-coefficient ranges:

(a) **High reflectivity:** $\Gamma_i \sim -\mathcal{U}(0.8, 0.9)$ (e.g., concrete, stone, glass). The fitted model uses $\Gamma = -0.85$.
(b) **Moderate reflectivity:** $\Gamma_i \sim -\mathcal{U}(0.6, 0.7)$ (e.g., plaster, painted drywall). The fitted model uses $\Gamma = -0.65$.
(c) **Low reflectivity:** $\Gamma_i \sim -\mathcal{U}(0.4, 0.5)$ (e.g., furnished offices, absorptive treatments). The fitted model uses $\Gamma = -0.45$.

For each reflectivity range, additive white Gaussian noise is added to achieve peak signal-to-noise ratios (PSNRs) $\{\infty, 40, 30, 25\}$ dB, where ∞ dB denotes the noise-free case. Noise is quantified using PSNR rather than classical SNR, following [45], as PSNR normalizes noise relative to the direct-path peak and thus provides a more conservative noise model that better reflects practical RIR conditions while enabling consistent comparison with prior work.

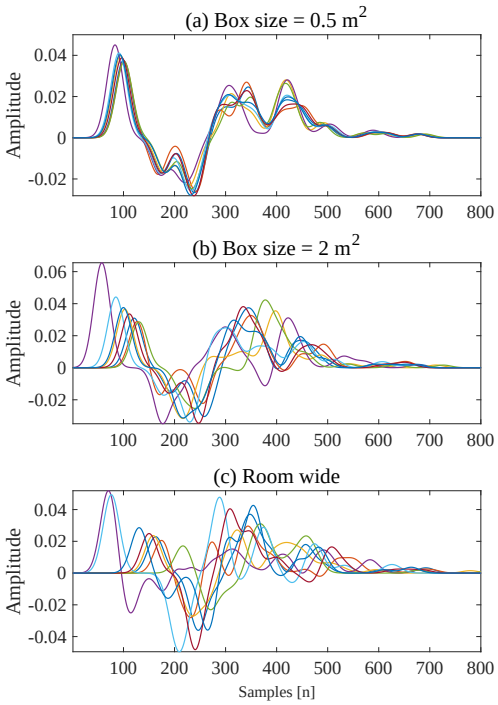

Fig. 4: Representative smoothed RIRs with $\sigma = 16$ for three microphone spreads: (a) $0.5\,\mathrm{m}^3$ cube, (b) $2\,\mathrm{m}^3$ cube, (c) distributed throughout the room.

D. Real-Data Experiments

We evaluate the proposed method using two sets of measured RIRs from [15] and [59]. In both cases, the employed RIR model incorporates reflections up to the third order as in Eq. (17), and all signals are sampled at 48 kHz. Other publicly available datasets were excluded due to the absence of accessible ground truth, the need for extensive preprocessing, or incompatible microphone/source configurations.

Preprocessing: Each measured RIR is first match-filtered with a sinc kernel to enhance early reflections. A coarse peak-picking stage then identifies the largest 100 peaks per RIR. Short temporal windows around these peaks are retained, while the remaining samples are treated as noise. This processing step does not require sophisticated signal processing or careful tuning. The resulting sparse RIRs are subsequently smoothed via convolution with a Gaussian kernel (Sec. III-A) to produce signals compatible with the model. The Gaussian kernel width is set to $\sigma = 16$ in both datasets. Although one might expect σ to scale inversely with the sampling frequency, this heuristic assumes noiseless, band-limited measurements and an exact forward model. In practice, higher-frequency RIR components are noise-dominated and poorly modeled; stronger relative smoothing therefore improves conditioning and convergence without sacrificing information relevant to the model. Fig. 5 shows an example of a measured RIR from [15] at successive processing stages alongside the corresponding model RIRs. Panel (a) shows the raw measured RIR, $\hat{\mathbf{h}}_m^{(\mathrm{raw})}$, overlaid with the model RIR $\mathbf{h}_m(\mathcal{Q})$ computed using Eq. (17). Panel (b) shows the RIR after match filtering with a sinc kernel and peak picking, also compared to $\mathbf{h}_m(\mathcal{Q})$. Panel (c) presents

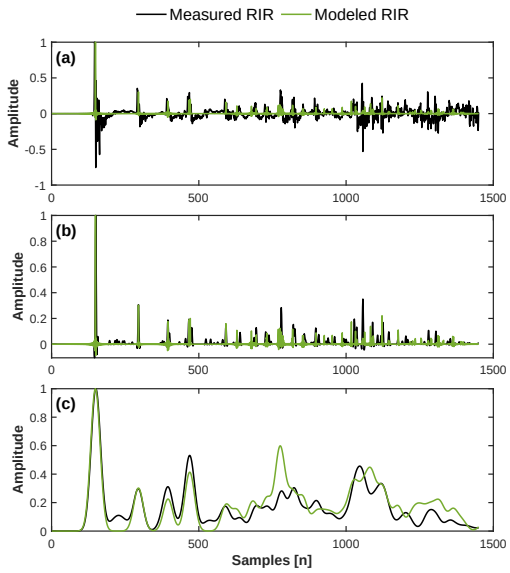

Fig. 5: Example of the processing of a measured RIR from [15] (a) measured $\hat{\mathbf{h}}_m^{(\text{raw})}$ compared to the model $\mathbf{h}_m(\mathcal{Q})$. (b) Processed measured RIR $\hat{\mathbf{h}}_m$ compared to the model $\mathbf{h}_m(\mathcal{Q})$. (c) Gaussian-smoothed $\hat{\mathbf{h}}_m^{(\sigma)}$ compared to the model $\mathbf{h}_m(\mathcal{Q};\sigma)$.

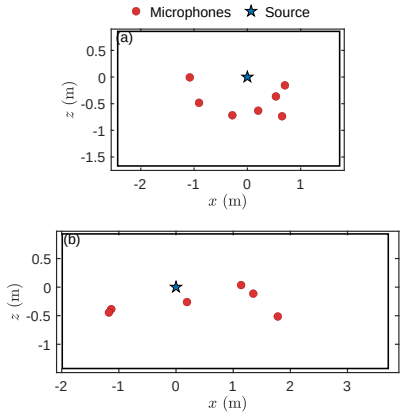

Fig. 6: Measurement setup showing xz view of microphones and source in (a) **Dataset 1** and (b) **Dataset 2**.

the Gaussian-smoothed processed RIR, $\hat{\mathbf{h}}_m^{(\sigma)}$, alongside the Gaussian kernel model RIR, $\mathbf{h}_m(\mathcal{Q};\sigma)$.

*1) Measured RIRs from [15] (**Dataset 1**):* This dataset contains RIRs measured at $M = 7$ microphones in a small, empty cuboid room measuring $5.97 \times 4.17 \times 2.53$ m. A single loudspeaker was used, with all microphones fixed in space. The measurement configuration used is illustrated in Fig. 6 (a). The room has concrete walls covered with 3 mm metal plates, resulting in a reverberation time of 1.49 s. Ground-truth room geometry was obtained using an infrared tracker (OptiTrack). The uniform reflection coefficient required by the model is estimated from the measured RIRs as described in Sec. III-A and is fixed at $\Gamma = 0.7$.

*2) The dEchorate room geometry example [59] (**Dataset 2**):* The second dataset uses the `data_rooge.mat` example from the dEchorate dataset. Although the full dataset contains more than six microphones, only the $M = 6$ microphones for a single source provided in the processed MATLAB example are used for ease of access. The recording setup is placed in a cuboid room measuring $6 \times 6 \times 2.4$ m. The measurement

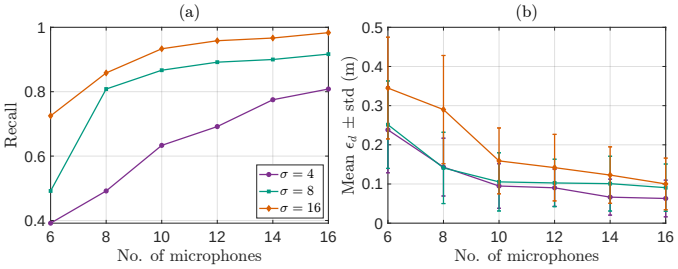

Fig. 7: **Experiment 1**. Performance vs. number of microphones M for Gaussian kernel widths $\sigma \in \{4, 8, 16\}$. (a) Recall of first-order image sources. (b) Mean Euclidean localization error ϵ_d. Whiskers indicate one standard deviation.

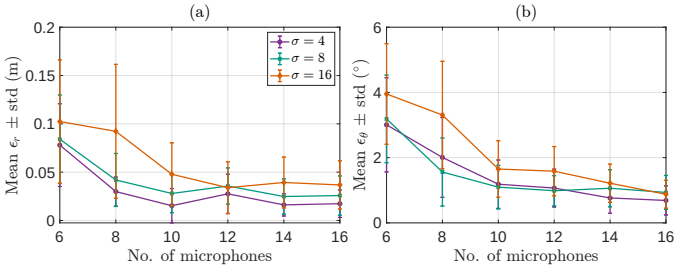

Fig. 8: **Experiment 1**. Performance vs. number of microphones M for $\sigma \in \{4, 8, 16\}$. (a) Mean radial error ϵ_r. (b) Mean angular error ϵ_θ. Whiskers indicate one standard deviation.

configuration is illustrated in Fig. 6 (b). All surfaces are covered with acoustic panels, allowing controlled reverberation times. In the chosen configuration (room 011111 as described in [59]), the floor is covered with hairy carpet (absorptive), while the ceiling and walls consist of 20 mm thick Formica and plaster panels (reflective). Ground-truth reflector positions are also available.

Since the floor reflection is one of the first two arriving reflections and has a lower estimated individual reflection coefficient ($\Gamma = 0.4$), while the other reflection has $\Gamma = 0.8$, the resulting average reflection coefficient ($\Gamma = 0.6$) is somewhat unrepresentative of the ground truth. Although a more accurate estimate could be obtained from the general decay of the RIR, we fix $\Gamma = 0.6$ to demonstrate that a moderate mismatch in the reflection coefficient does not completely disrupt the algorithm's performance.

VI. Results and Discussion

This section reports the results from the three simulation studies and the two real-data evaluations introduced in Sec. V.

A. Simulation Results

Experiment 1: Figs. 7 and 8 summarize performance as a function of the number of microphones $M \in \{6, 8, 10, 12, 14, 16\}$ and Gaussian kernel widths $\sigma \in \{4, 8, 16\}$. Fig. 7(a) shows recall of first-order image sources, while Fig. 7(b) reports mean Euclidean localization error ϵ_d. Fig. 8 further decomposes the error into radial (ϵ_r) and angular (ϵ_θ) components. It can be observed that recall increases steadily with M for all kernel widths. For $\sigma = 16$, recall rises from about 73% at $M = 6$ to 98% at $M = 16$, while for $\sigma = 4$, recall goes from 39% to 81%. The largest gains occur between $M = 6$ and $M = 10$, with diminishing returns

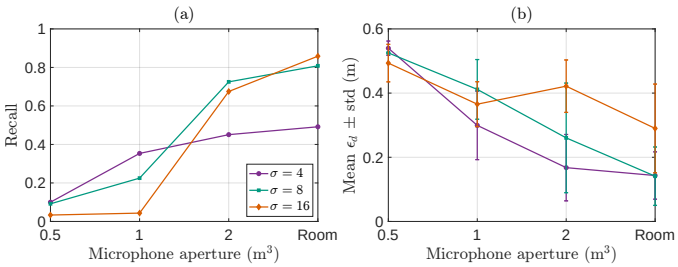

Fig. 9: **Experiment 2**. Performance vs. microphone spacing (box size) for Gaussian kernel widths $\sigma \in \{4, 8, 16\}$. (a) Recall of first-order image sources. (b) Mean Euclidean localization error ϵ_d. Whiskers indicate one standard deviation.

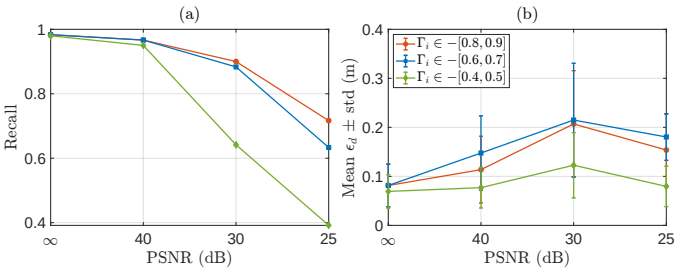

Fig. 10: **Experiment 3**. Performance as a function of PSNR (clean, 40, 30, 25 dB) for three reflectivity ranges (high, moderate, low) and $M = 16$ microphones. (a) Recall of first-order image sources. (b) Mean Euclidean localization error with standard-deviation whiskers, computed over correctly retrieved first-order image sources only.

beyond $M \approx 12$ (e.g., for $\sigma = 16$, only $\approx 5\%$ improvement from $M = 12$ to $M = 16$). Localization accuracy shows a similar trend. For $\sigma = 16$, mean Euclidean error decreases from roughly $0.35\,\mathrm{m}$ at $M = 6$ to $0.10\,\mathrm{m}$ at $M = 16$. Radial and angular errors also decrease substantially, confirming that additional microphones improve both direction and distance estimation. For a fixed M, increasing σ improves recall. At $M = 10$, recall rises from $\approx 63\%$ ($\sigma = 4$) to 87% ($\sigma = 8$) and 93% ($\sigma = 16$). The effect of σ on localization depends on M: with few microphones, narrower kernels preserve sharper timing information and improve Euclidean error (e.g., at $M = 6$, ϵ_d ranges from $0.24\,\mathrm{m}$ for $\sigma = 4$ to $0.35\,\mathrm{m}$ for $\sigma = 16$), whereas for $M = 16$, errors converge ($\approx 0.06 - 0.10\,\mathrm{m}$) and become less sensitive to σ. In summary, increasing M improves both recall and localization, with diminishing returns beyond ≈ 12 microphones. Larger σ enhances robustness and recall, while smaller σ provides sharper geometric precision under low-M conditions. Deploying $M \geq 8$ microphones with $\sigma = 8$–16 balances reliability and accuracy, giving recall $> 80\%$ and $\epsilon_d < 0.03\,\mathrm{m}$.

Experiment 2: Fig. 9 shows the effect of microphone spatial spread on recall and mean Euclidean error for $M = 8$, with Gaussian kernel widths $\sigma \in \{4, 8, 16\}$. For the smallest aperture ($0.5\,\mathrm{m}^3$ cube), recall is low for all σ (3.3% for $\sigma = 16$, 10% for $\sigma = 4$) and mean Euclidean error exceeds $0.49\,\mathrm{m}$ for $\sigma = 16$. In this configuration, the smoothed RIRs are highly similar across channels (Fig. 4a). With so little variation, the optimization problem becomes poorly conditioned, producing noisy and poorly informative gradients that limit the effectiveness of the optimization and result in low recall. These results demonstrate that tightly clustered microphones provide insufficient spatial diversity to reliably resolve early reflections, and reducing σ only partially improves performance by sharpening the smoothed RIR peaks. Increasing the aperture improves performance. In a $2\,\mathrm{m}^3$ cube, recall reaches 67.5% for $\sigma = 16$ with mean Euclidean error $0.42\,\mathrm{m}$. Full-room placements give the highest recall (85.8%) and lowest Euclidean error ($0.29 \pm 0.14\,\mathrm{m}$). The smoothing width σ mainly affects performance when spatial diversity is limited. In the $1\,\mathrm{m}^3$ configuration, reducing σ from 16 to 4 increases recall from 4.3% to 35.3% and lowers Euclidean error from $0.37\,\mathrm{m}$ to $0.30\,\mathrm{m}$. Overall, these results indicate that microphone geometry is an important factor affecting performance. Increasing array aperture consistently improves recall and localization, while smaller σ can partially compensate for limited spatial diversity by increasing temporal resolution.

Experiment 3: Fig. 10 shows recall and the mean Euclidean error ϵ_d as a function of PSNR for three reflectivity ranges (high, moderate, and low) and $M = 16$ microphones. As expected, recall decreases as PSNR is reduced from noiseless conditions to 40, 30, and 25 dB. The degradation is modest down to $30\,\mathrm{dB}$ but becomes substantial at $25\,\mathrm{dB}$, where weaker reflections approach the noise floor and become difficult to detect, an effect amplified by the Gaussian smoothing. Reflectivity further compounds this behavior. Lower reflectivity reduces reflection amplitudes and increases mismatch with the uniform-Γ model used during fitting, slightly distorting the objective. In noiseless and $40\,\mathrm{dB}$ settings, recall is similar across reflectivity levels; however, differences emerge at $30\,\mathrm{dB}$ and become pronounced at $25\,\mathrm{dB}$, where low-reflectivity cases exhibit the strongest decline. Despite this degradation at low reflectivity and $25\,\mathrm{dB}$ PSNR, the method still retrieves a nontrivial fraction of first-order image sources (approximately 40%), indicating notable robustness. Localization accuracy on the correctly retrieved image sources is less affected by noise. The mean Euclidean error ϵ_d increases gradually down to $30\,\mathrm{dB}$ but remains relatively stable overall. The decrease observed at $25\,\mathrm{dB}$ is due to selection bias: with many weak reflections undetected, the error is computed only on the few strong early arrivals that remain reliable at low PSNR. These results illustrate that when reflections remain above the noise floor, their estimated positions remain comparatively accurate even under low reflectivity.

B. Real-data Results

Dataset 1 [15]: Table II reports per-image and average metrics across $M = 7$ microphones. The method successfully detects five out of six first-order images, achieving a recall of 83.33%. The mean Euclidean, radial, and angular errors for correctly identified images are $0.37\,\mathrm{m}$, $0.10\,\mathrm{m}$, and $5.60°$, respectively. These outcomes closely align with the simulation trends: for a comparable setting with $M = 8$, simulations yielded average recall $\approx 86\%$ and $\epsilon_d \approx 0.29\,\mathrm{m}$. The small gap is consistent with the one-microphone difference and real-room effects (non-uniform reflectivities, local SNR variations, non-specular reflections and scattering). Individual Euclidean errors range from $0.11\,\mathrm{m}$ to $0.59\,\mathrm{m}$, reflecting the influence of local SNR variations and wall-specific amplitude mismatches. Even with a single uniform reflection coefficient ($\Gamma = 0.7$),

TABLE II: Per first-order-image source error and detection metrics for **Dataset 1**

Image source	Euclid. (m)	Rad. (m)	Ang. (°)	Correct
1 (floor)	0.13	0.05	2.19	true
2 (wall)	3.52	3.44	6.41	false
3 (wall)	0.50	0.09	8.28	true
4 (wall)	0.51	0.16	7.37	true
5 (wall)	0.59	0.17	6.58	true
6 (ceiling)	0.11	0.01	3.57	true
Average:	**0.37**	**0.10**	**5.60**	**Recall: 83.33%**

TABLE III: Per first-order-image source error and detection metrics for **Dataset 2**.

Image source	Euclid. (m)	Rad. (m)	Ang. (°)	Correct
1 (wall)	0.74	0.41	9.32	true
2 (wall)	0.86	0.17	6.60	false
3 (wall)	0.82	0.47	4.20	true
4 (wall)	1.82	1.74	8.00	false
5 (floor)	0.29	0.11	5.51	true
6 (ceiling)	0.03	0.01	0.93	true
Average:	**0.47**	**0.25**	**4.99**	**Recall: 66.67%**

the algorithm robustly recovers most first-order image sources. The mean Euclidean error corresponds to a TOA deviation of approximately $1.08\,\text{ms}$, or 52 samples at $48\,\text{kHz}$, demonstrating that the method maintains decent temporal precision despite Gaussian model approximations.

***Dataset 2** [59]:* Table III reports per-image and average metrics across $M = 6$ microphones. The method achieves a recall of 66.67%, successfully identifying four out of six first-order images, including the absorptive floor. The mean Euclidean, radial, and angular errors for correct estimates are $\epsilon_d = 0.47\,\text{m}$, $\epsilon_r = 0.25\,\text{m}$, and $\epsilon_\theta = 4.99°$. These results are also consistent with simulations at similar microphone counts: for $M = 6$, simulations reported recall $\approx 73\%$ and $\epsilon_d \approx 0.35\,\text{m}$. As expected, the reduced number of microphones and non-uniform surface reflectivities slightly degrade performance compared with **Dataset 1**. In particular, amplitude mismatches due to the single Γ assumption primarily affect the signal fit but have limited impact on geometric recovery. Highly absorptive surfaces reduce SNR for corresponding reflections, but sufficient channel diversity allows most early arrivals to be detected. Overall, real-data performance aligns well with the simulations and demonstrates robustness under practical conditions. **Dataset 1** is consistent with the $M \approx 8$ trend, showing high recall and reasonable error, while **Dataset 2** exhibits the expected degradation with fewer microphones and non-uniform absorption.

VII. Conclusion

We presented an RIR-level inversion method for recovering first-order image-source positions and thereby the geometry of the room. The method parametrizes the entire early-reflection structure (up to third order) in terms of first-order image sources and fits a differentiable ISM-based forward model without TOA/TDOA/DOA extraction or echo labeling. A cosine-similarity objective improves robustness to amplitude scaling and other unknowns, while replacing bandlimited sinc kernels with localized Gaussian kernels yields smooth gradients and a better-conditioned optimization landscape. We further proposed a practical multi-stage algorithm that combines multi-start Adam optimization with clustering to stabilize and refine solutions across stages. Extensive simulations show that: (i) increasing the number of microphones and the array's spatial aperture consistently improves recall and accuracy; (ii) larger Gaussian widths produce smoother objectives and higher recall at the cost of slightly reduced best-case localization accuracy; and (iii) performance degrades gracefully under moderate sensor noise and lower wall reflectivity. Experiments on two measured-RIR datasets recover a majority of first-order image sources, including under moderate modeling mismatch, underscoring practical viability. While these results are encouraging, several design choices entail trade-offs. Gaussian smoothing, though beneficial for optimization, can broaden and occasionally merge closely spaced arrivals, limiting ultimate localization accuracy. The use of a single broadband reflection coefficient introduces amplitude mismatch when surface properties vary. Moreover, compact microphone layouts reduce inter-channel diversity and can impair identifiability. Refinement of the proposed algorithm, as well as exploration of alternative optimization strategies, could be pursued in future work to improve reliability under non-ideal conditions and to achieve more accurate image-source retrieval. Overall, the proposed formulation offers a simple and robust route to room-geometry inference, providing a strong foundation for additional boundary-parameter estimation.

References

[1] D. Zotkin, R. Duraiswami, and L. Davis, "Rendering localized spatial audio in a virtual auditory space," *IEEE Trans. Multimedia*, vol. 6, no. 4, pp. 553–564, 2004.

[2] F. Keyrouz and K. Diepold, "Binaural source localization and spatial audio reproduction for telepresence applications," *Presence*, vol. 16, no. 5, pp. 509–522, 2007.

[3] H. Kim, L. Remaggi, P. J. B. Jackson, and A. Hilton, "Immersive spatial audio reproduction for vr/ar using room acoustic modelling from 360 images," in *IEEE Conference on Virtual Reality and 3D User Interfaces (IEEE VR)*, 2019.

[4] A. S. Bregman, "Auditory scene analysis: The perceptual organization of sound," 1990.

[5] J. Ahrens, "Analytic methods of sound field synthesis," 2012.

[6] P. A. Naylor and N. D. Gaubitch, "Speech dereverberation," 2010.

[7] J. Benesty, T. Gänsler, D. R. Morgan, M. M. Sondhi, S. L. Gay, *et al.*, "Advances in network and acoustic echo cancellation," 2001.

[8] C. Evers, H. W. Löllmann, H. Mellmann, A. Schmidt, H. Barfuss, P. A. Naylor, and W. Kellermann, "The LOCATA challenge: Acoustic source localization and tracking," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 1620–1643, 2020.

[9] M. A. Poletti, "Active acoustic systems for the control of room acoustics," 2013.

[10] S. Siltanen, T. Lokki, L. Savioja, and C. Lynge Christensen, "Geometry reduction in room acoustics modeling," *Acta Acustica united with Acustica*, vol. 94, no. 3, pp. 410–418, 2008.

[11] A. H. Moore, M. Brookes, and P. A. Naylor, "Room geometry estimation from a single channel acoustic impulse response," in *Proc. 21st European Signal Process. Conf. (EUSIPCO '13)*, pp. 1–5, 2013.

[12] F. Antonacci, J. Filos, M. R. Thomas, E. A. Habets, A. Sarti, P. A. Naylor, and S. Tubaro, "Inference of room geometry from acoustic impulse responses," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 20, no. 10, pp. 2683–2695, 2012.

[13] I. Dokmanić, R. Parhizkar, A. Walther, Y. M. Lu, and M. Vetterli, "Acoustic echoes reveal room shape," *Proceedings of the National Academy of Sciences*, vol. 110, no. 30, pp. 12186–12191, 2013.

[14] M. Crocco, A. Trucco, and A. Del Bue, "Uncalibrated 3D room geometry estimation from sound impulse responses," *Journal of the Franklin Institute*, vol. 354, no. 18, pp. 8678–8709, 2017.

[15] S. Park and J.-W. Choi, "Iterative echo labeling algorithm with convex hull expansion for room geometry estimation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 1463–1478, 2021.

[16] Y. El Baba, A. Walther, and E. A. Habets, "3D room geometry inference based on room impulse response stacks," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 5, pp. 857–872, 2017.

[17] L. Remaggi, P. J. Jackson, P. Coleman, and W. Wang, "Room boundary estimation from acoustic room impulse responses," in *2014 Sensor Signal Processing for Defence (SSPD)*, pp. 1–5, IEEE, 2014.

[18] T. Rajapaksha, X. Qiu, E. Cheng, and I. Burnett, "Geometrical room geometry estimation from room impulse responses," in *Proc. 2016 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '16)*, pp. 331–335, IEEE, 2016.

[19] M. Lovedee-Turner and D. Murphy, "Three-dimensional reflector localisation and room geometry estimation using a spherical microphone array," *J. Acoust. Soc. Amer.*, vol. 146, no. 5, pp. 3339–3352, 2019.

[20] F. Antonacci, A. Sarti, and S. Tubaro, "Geometric reconstruction of the environment from its response to multiple acoustic emissions," in *Proc. 2010 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '10)*, pp. 2822–2825, IEEE, 2010.

[21] A. Canclini, F. Antonacci, M. R. Thomas, J. Filos, A. Sarti, P. A. Naylor, and S. Tubaro, "Exact localization of acoustic reflectors from quadratic constraints," in *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 17–20, IEEE, 2011.

[22] E. Nastasia, F. Antonacci, A. Sarti, and S. Tubaro, "Localization of planar acoustic reflectors through emission of controlled stimuli," in *Proc. 19th European Signal Process. Conf. (EUSIPCO '11)*, pp. 156–160, IEEE, 2011.

[23] I. Jager, R. Heusdens, and N. D. Gaubitch, "Room geometry estimation from acoustic echoes using graph-based echo labeling," in *Proc. 2016 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '16)*, pp. 1–5, IEEE, 2016.

[24] M. Coutino, M. B. Møller, J. K. Nielsen, and R. Heusdens, "Greedy alternative for room geometry estimation from acoustic echoes: A subspace-based method," in *Proc. 2017 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '17)*, pp. 366–370, IEEE, 2017.

[25] I. Dokmanić, Y. M. Lu, and M. Vetterli, "Can one hear the shape of a room: The 2-d polygonal case," in *Proc. 2011 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '11)*, pp. 321–324, 2011.

[26] J. Filos, E. A. Habets, and P. A. Naylor, "A two-step approach to blindly infer room geometries," in *Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC), Tel Aviv, Israel*, Citeseer, 2010.

[27] M. Kreković, I. Dokmanić, and M. Vetterli, "Echoslam: Simultaneous localization and mapping with acoustic echoes," in *Proc. 2016 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '16)*, pp. 11–15, 2016.

[28] K. Kowalczyk, E. A. Habets, W. Kellermann, and P. A. Naylor, "Blind system identification using sparse learning for TDOA estimation of room reflections," *IEEE Signal Processing Letters*, vol. 20, no. 7, pp. 653–656, 2013.

[29] E. Mabande, K. Kowalczyk, H. Sun, and W. Kellermann, "Room geometry inference based on spherical microphone array eigenbeam processing," *J. Acoust. Soc. Amer.*, vol. 134, no. 4, pp. 2773–2789, 2013.

[30] S. Tervo and A. Politis, "Direction of arrival estimation of reflections from room impulse responses using a spherical microphone array," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 10, pp. 1539–1551, 2015.

[31] T. Shlomo and B. Rafaely, "Blind amplitude estimation of early room reflections using alternating least squares," in *Proc. 2021 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '21)*, pp. 476–480, 2021.

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

[33] K. MacWilliam, F. Elvander, and T. van Waterschoot, "Simultaneous acoustic echo sorting and 3-D room geometry inference," in *Proc. 2023 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '23)*, pp. 1–5, IEEE, 2023.

[34] R. Fernando, R. Rameshbabu, K. Sarkar, A. C. Singer, and R. M. Corey, "Neural particle on a sphere to estimate image source locations for room impulse response interpolation," 2025.

[35] L. Remaggi, H. Kim, P. J. Jackson, F. M. Fazi, and A. Hilton, "Acoustic reflector localization and classification," in *Proc. 2018 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '18)*, pp. 201–205, IEEE, 2018.

[36] S. Tervo and T. Tossavainen, "3D room geometry estimation from measured impulse responses," in *Proc. 2012 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '12)*, pp. 513–516, IEEE, 2012.

[37] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: An overview," *EURASIP Journal on Advances in Signal Processing*, vol. 2006, pp. 1–19, 2006.

[38] I. J. Kelly and F. M. Boland, "Detecting arrivals in room impulse responses with dynamic time warping," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 22, no. 7, pp. 1139–1147, 2014.

[39] B. L. Sturm and G. Defrance, "Detection and estimation of arrivals in room impulse responses by greedy sparse approximation," in *Proc. 18th European Signal Process. Conf. (EUSIPCO '10)*, pp. 1934–1938, IEEE, 2010.

[40] H. Sun, E. Mabande, K. Kowalczyk, and W. Kellermann, "Localization of distinct reflections in rooms using spherical microphone array eigenbeam processing," *J. Acoust. Soc. Amer.*, vol. 131, no. 4, pp. 2828–2840, 2012.

[41] S. Damiano, F. Borra, A. Bernardini, F. Antonacci, and A. Sarti, "Soundfield reconstruction in reverberant rooms based on compressive sensing and image-source models of early reflections," in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 366–370, 2021.

[42] Y. Chi, L. L. Scharf, A. Pezeshki, and A. R. Calderbank, "Sensitivity to basis mismatch in compressed sensing," *IEEE Trans. Signal Process.*, vol. 59, no. 5, pp. 2182–2195, 2011.

[43] T. Sprunck, A. Deleforge, Y. Privat, and C. Foy, "Gridless 3d recovery of image sources from room impulse responses," *IEEE signal processing letters*, vol. 29, pp. 2427–2431, 2022.

[44] F. Ribeiro, D. Florencio, D. Ba, and C. Zhang, "Geometrically constrained room modeling with compact microphone arrays," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 20, no. 5, pp. 1449–1460, 2012.

[45] T. Sprunck, A. Deleforge, Y. Privat, and C. Foy, "Fully reversing the shoebox image source method: From impulse responses to room parameters," *IEEE Trans. Audio Speech Lang. Process.*, 2025.

[46] T. van Waterschoot, "Deep, data-driven modeling of room acoustics: literature review and research perspectives," in *Proc. 11th Convent. Europ. Acoust. Assoc. Forum Acusticum/EuroNoise 2025*, (Málaga, Spain), pp. 4065–4072, June 2025.

[47] W. Yu and W. B. Kleijn, "Room acoustical parameter estimation from room impulse responses using deep neural networks," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 436–447, 2020.

[48] W. Yu and W. B. Kleijn, "Estimation of source and receiver positions, room geometry and reflection coefficients from a single room impulse response," *arXiv preprint arXiv:2301.09198*, 2023.

[49] I. Yeon and J.-W. Choi, "3d room geometry inference from multichannel room impulse response using deep neural network," *arXiv preprint arXiv:2401.10453*, 2024.

[50] H. N. Bicer, C. Tuna, A. Walther, and E. A. Habets, "Data-driven joint detection and localization of acoustic reflectors," in *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSP'24)*, pp. 745–749, IEEE, 2024.

[51] C. Tuna, A. Akat, H. N. Bicer, A. Walther, and E. A. Habets, "Data-driven 3D room geometry inference with a linear loudspeaker array and a single microphone," *arXiv preprint arXiv:2308.14611*, 2023.

[52] J. B. Keller, "The scope of the image method," *Communications on pure and applied mathematics*, vol. 6, no. 4, pp. 505–512, 1953.

[53] A. Deleforge, *Can One Hear the Walls of a Room? Physics- and Data-Driven Inverse Methods for Acoustic Signal Processing*. Accreditation to supervise research, Université de Strasbourg, Nov. 2025.

[54] R. F. Hoskins, "Delta functions: Introduction to generalised functions," 2009.

[55] A. S. Householder, "Unitary triangularization of a nonsymmetric matrix," *Journal of the ACM (JACM)*, vol. 5, no. 4, pp. 339–342, 1958.

[56] M. R. Schroeder, "New method of measuring reverberation time," *The Journal of the Acoustical Society of America*, vol. 37, no. 3, pp. 409–412, 1965.

[57] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.

[58] K. Khan, S. U. Rehman, K. Aziz, S. Fong, and S. Sarasvady, "Dbscan: Past, present and future," in *The fifth international conference on the applications of digital information and web technologies (ICADIWT 2014)*, pp. 232–238, IEEE, 2014.

[59] D. Di Carlo, P. Tandeitnik, C. Foy, N. Bertin, A. Deleforge, and S. Gannot, "dEchorate: a Calibrated Room Impulse Response Dataset for Echo-aware Signal Processing," *EURASIP Journal on Audio, Speech, and Music Processing*, Nov. 2021.