Fully Reversing the Shoebox Image Source Method: From Impulse Responses to Room Parameters

Tom Sprunck, Antoine Deleforge, Yannick Privat, Cédric Foy

Fully Reversing the Shoebox Image Source Method: From Impulse Responses to Room Parameters

Tom Sprunck, Antoine Deleforge, Yannick Privat and Cédric Foy

***Abstract*—We present an algorithm that fully reverses the shoebox image source method (ISM), a popular and widely used room impulse response (RIR) simulator for cuboid rooms introduced by Allen and Berkley in 1979. More precisely, given a discrete multichannel RIR generated by the shoebox ISM for a microphone array of known geometry, the algorithm reliably recovers the 18 input parameters. These are the 3D source position, the 3 dimensions of the room, the 6-degrees-of-freedom room translation and orientation, and an absorption coefficient for each of the 6 room boundaries. The approach builds on a recently proposed gridless image source localization technique combined with new procedures for room axes recovery and first-order-reflection identification. Extensive simulated experiments reveal that near-exact recovery of all parameters is achieved for a 32-element, 8.4-cm-wide spherical microphone array and a sampling rate of 16 kHz using fully randomized input parameters within rooms of size $2\times2\times2$ to $10\times10\times5$ meters. Estimation errors decay towards zero when increasing the array size and sampling rate. The method is also shown to strongly outperform a known baseline, and its ability to extrapolate RIRs at new positions is demonstrated. Crucially, the approach is strictly limited to low-passed discrete RIRs simulated using the vanilla shoebox ISM. Nonetheless, it represents to our knowledge the first algorithmic demonstration that this difficult inverse problem is in-principle fully solvable over a wide range of configurations.**

***Index Terms*—Room Shape, Acoustics, Room Impulse Response, Sound Field, Reflectors, Echoes, Image Source, Gridless**

I. Introduction

HEARING the shape of a room, or more formally the problem of recovering the properties of a room boundary from the acoustic measurements of one or several sound sources inside of it, is a difficult inverse problem that has intrigued researchers in audio signal processing and room acoustics for many years. Beyond its folklore nature, solutions to this problem could benefit applications in augmented reality [1], [2], room compensation [3], sound field reconstruction [4], [5], robotic navigation [6], [7], or room acoustic diagnosis [8].

A steady number of approaches have been proposed to tackle different facets of this question along the past two decades [8]–[50], but their direct comparison is nearly impossible due to the many variations of the problem that have been considered. Most approaches make use of room impulse responses (RIRs), but some tackle the question *blindly*, namely, with no knowledge of the source signals [12]–[14], [20], [24], [33], [41], [45], [46], while other approaches assume that the times of arrival of acoustic reflections are directly available [8], [15], [16], [22], [26], [27], [35], [36]. Most approaches make use of multiple microphones at arbitrary known locations, but some use a single microphone [11], [15], [26], [27], [29], [42], [43], [48] or specific microphone array geometries such as spherical [20], [23], [31], [40], [45], [46], circular [14], [25], [50], linear [9], [28], or others [21], [30], [39]. Most approaches use a single sound source, but some use multiple sources [8], [11], [12], [16], [21], [30], [33], [35], [37], [39], [40] or a linear loudspeaker array [38], [43], [48]. Some approaches require specific geometrical assumptions on the setup, such as a rectangular or cuboid ("shoebox") room [21], [26], [33], [42], [47], [48] or having all sources and microphones lying on a plane parallel to both the floor and ceiling [11], [19], [21], [43], [48]–[50]. Some approaches focus on the 2D case [13]–[15], [17], [26]–[28], [33], [44] or the 1D case [29]. The question of what is to be recovered also varies greatly. The distance and orientation of one acoustic reflector [12], [14], [17], [30], several reflectors [11], [16], [22], [23], [35], [37], [39], [50] or all of the reflectors [13], [15], [18], [19], [21], [26], [29], [36], [38], [40], [43], [48], [49] in the room? The unlabelled time differences of arrival [24], [34], [41], directions of arrival [20], [25], [28], [31], or 3D positions [46], [47] of image sources? Additional properties of the reflectors such as their absorption [8], [10], [27], [32], [33], [42], [44], [45], [47] or their size [39]? Finally, the considered noise, sensor, and sound propagation models may differ widely across existing approaches.

A number of successes have been obtained over the years, including demonstrations on real measured acoustic data [18], [20], [22], [25], [37], [39], [40], [42], [48]. However, because there seems to be nearly as many ways of framing the question as there are research articles on the topic, the more fundamental question of **whether the problem is solvable at all for a clearly specified and broad enough set of assumptions** remains largely open to date. In this article, we do not introduce a solution that is readily applicable to real data, but turn our attention towards this more fundamental question instead. To this aim, we focus on a simple but fully specified forward room acoustic model, namely, the well-known shoebox image source method (ISM) proposed by Allen and Berkley in 1979 [51]. We then frame the question as one of *algorithmic reversibility*, namely, is there an algorithm that, given the output of the shoebox ISM, can reliable recover all of its input? More precisely, given a discrete, ideally

This work was made with the support of the French National Research Agency through project DENISE (ANR-20-CE48-0013).

Antoine Deleforge and Tom Sprunck are with IRMA, CNRS, Université de Strasbourg, Inria, 67000 Strasbourg, France.

Yannick Privat is with IECL, Université de Lorraine, CNRS, Inria, BP 70239 54506 Vandœuvre-lès-Nancy Cedex, France (yannick.privat@univ-lorraine.fr)

Yannick Privat is with Institut Universitaire de France (IUF).

Cédric Foy is with UMRAE, Cerema, Univ. Gustave Eiffel, Ifsttar, Strasbourg, 67035, France (cedric.foy@cerema.fr).

low-passed, multichannel RIR generated by the ISM for a microphone array of known geometry, we ask whether the following 18 parameters can be recovered:

- The 6-degrees-of-freedom translation and orientation of the room in the microphone array coordinate frame;
- The 3-dimensional source position in the microphone array coordinate frame;
- The 3 dimensions of the room;
- Absorption coefficients for the 6 room surfaces.

We provide an open-source algorithm and extensive experimental results that suggest that the answer to this question is *yes*, under a broad range of randomized input parameters, for sufficiently large microphone arrays and sufficiently high frequencies of sampling. In particular, near exact inversion is achieved, with geometrical errors in the order of millimeters and hundredth of degrees across hundreds of randomly generated rooms of size $2\times2\times2$ to $10\times10\times5$ meters, using a 32-element spherical microphone array of diameter 8.4 cm and a frequency of sampling of 16 kHz. These errors keep steadily decreasing when increasing the array size and sampling rate. Errors below 3 mm are also obtained in 95% of our test cases for an 8-element non-spherical array, outperforming by an order of magnitude the well-known baseline of Dokmanic et al. [22] to which oracle times of arrival are provided, at a fraction of the computational cost. We finally show that the parameters estimated by the proposed inverse algorithm can be fed back to the forward model to *extrapolate* RIRs at any source-array placements in the room, with signal-to-error ratios above 20 dB for large enough arrays. RIR interpolation has been recently investigated in, *e.g.*, [4], [5].

The presented algorithm builds on a recently proposed method by the authors that estimates a 3D image-source *point cloud* up to a given range from a discrete multichannel RIR [47]. We devise here a new three-stage procedure that recovers the 18 parameters of interest from such a point cloud. First, the 3D room orientation is estimated. Second, the true source and first order image sources are labeled based on this orientation. Third, the remaining parameters are estimated based on the locations and amplitudes of labeled image sources.

The remainder of this article is organized as follows. Section II recalls relevant background and offers a review of the state of the art. Section III reviews the shoebox image-source forward model and the image source localization procedure used in this study. Section IV describes the proposed room parameter recovery algorithm. Section V presents extensive simulated experiments and results supporting the algorithmic reversibility claim. Finally, we provide concluding remarks and perspectives in Section VI.

II. Background and State of the Art

The key physical phenomenon making room geometry estimation from audio measurements possible at all is that of *early acoustic reflections*. When sound propagates from a source inside of a room, it is reflected on surfaces before reaching microphones. This materializes into delayed and filtered copies of the emitted signal inside the measured time-domain signals, that are commonly referred to as *echoes*. The *time of arrival* (TOA) of an echo at a microphone is proportional to the length of the corresponding reflected propagation path, while the *time differences of arrival* (TDOAs) of an echo between two or more microphones are linked to the *direction of arrival* (DOA) of the corresponding reflected propagation path. The core idea of nearly all existing methods in the field is to estimate such quantities from measured signals, to prune, sort and label echoes, and to solve for the acoustic-scene geometry based on the recovered information. A litterature review of the works tackling some or all of these steps is proposed in the remainder of this section.

The reflector associated to the TOA of a first-order propagation path from a source to a microphone is known to be tangential to an ellipsoid whose focci are the corresponding source and microphone positions. Assuming the latter are known, a number of early approaches, referred to as *direct localization* in [37], have hence focused on detecting, pruning, clustering and localizing tangent lines to multiple ellipses in the 2D case [11], [13], [14], [17], [19], [21], or tangent planes to multiple ellipsoids in the 3D case [16], [30], [37]. An alternative to this is to combine the TOAs and DOAs of echoes to obtain the 3D locations of their associated image sources. Reflectors can then be localized as the bi-secting planes between a true source and its first order image sources, as in [15], [18], [22], [23], [36], [37], [39], [40], [43]. This approach is referred to as *image source reversion* in [37] and is the one employed in this article.

Several early works in the field assume that TOAs are trivial to estimate from room impulse responses (RIRs) using peak picking [11], [19] or consider them readily available [15], [16], [22], [26], [27], [35], [36]. This would be the case if microphones, sources and reflectors had perfectly flat responses up to very high frequencies, but this is never true in practice. This band-limitedness results in a significant *smearing* of echoes, blurring the location of their peaks and making them overlap and interfere with each other in the time domain. An analogous phenomenon occurs in the 1D and 2D DOA domains, and is reinforced by the limited diameter of microphone arrays. Interference is all the more present since echoes are, by definition, strongly correlated with each other and with direct-path signals. Due to this, the tasks of TOA, TDOA and DOA estimation of early acoustic reflections has been the focus of significant research effort. The vast majority of existing techniques proceed by some form of peak-picking over a *discretized* time domain [12], [13], [17], [21], [23], [24], [30], [34], [37], [38], DOA domain [14], [20], [30], [40], [46], joint TOA-DOA domain [25], [39], [43], 3D space [18] or ray space [9], [28]. To improve the separation and sharpness of objects inside such discrete grids, some methods leverage sparsity-based techniques [18], [24], [33], [34], [43], [46] or ad-hoc image processing tools [25], [28], [39], [40]. Despite these efforts, operating over discrete time or space suffers from intrinsic limitations. First, the separability of peaks is fundamentally limited. This has led many authors to impose additional constraints on the geometrical setup to ensure separation. Second, for 3D image-source localization, the required discrete-grid size grows cubically in the desired range and precision. This fundamentally limits the achievable

resolution under reasonable computational constraints. For instance, in [18], a sparse problem over a 3D grid of 900k points needs to be solved to achieve an angular resolution of $\approx 4^\circ$ and a distance resolution of ≈ 2 cm, while restricting the array-reflector distances to at most 3.5 m. Third, sparse optimization over a discrete grid fundamentally suffers from the so-called *basis-mismatch* problem [52], [53], generally requiring the use of ad-hoc post-processing steps.

There are a few notable exceptions to this discrete grid-search paradigm [27], [29], [31], [41], [47]. In [27], a class of 2D room geometries is selected (rectangle, L-shaped) and the continuous shape dimensions are directly optimized by minimizing a distance between measured and image-source TOAs, using a genetic algorithm. In [29], the wall- and source-to-wall distances in a 1D room are continuously optimized based on resonant frequencies. In [31], non-linear minimization of a likelihood-based cost function in the spherical harmonics domain is utilized to jointly estimate the continuous DOA of a fixed number of reflectors. In [41], the TDOAs of echoes are blindly estimated in the continuous time domain by leveraging an infinite-dimensional convex relaxation of the problem and the *sliding Frank-Wolfe* algorithm [52]. In [47], a similar approach is employed in 3D space to recover the continuous 3D positions of all image sources within a given range from a multichannel RIR. The present work builds on this last approach.

Many of the above-reviewed methods estimate TOAs and/or TDOAs independently across individual channels and/or channel pairs. To leverage these quantities for geometry estimation, they need to be associated to reflectors, a procedure referred to as *echo sorting*. This difficult combinatorial problem is the focus of [15], [22], [35]. The need for echo sorting is bypassed by methods that directly localize image sources from RIRs [18], [23], [31], [37], [39], [40], [43], [46], [47], such as the one employed in this work.

Once image sources are localized, a necessary subsequent step is to *label* them, namely, identify their order of reflection. In the literature, labeling is typically performed by ad-hoc algorithms that exploit the geometrical constraints at hands, *e.g.*, [18], [22], [36], [40]. They are often tailored to the specific class of source-microphone-room setup under consideration, and may hence be hard to generalize. In this work, a new approach to labeling is presented. We leverage the fact that the recently proposed image-source localization method in [47] can recover a much larger number of image sources than previously possible. We present a new technique that robustly estimate the 3D *orientation* of the room in the microphone array frame from such image-source point cloud. This specific task has not been investigated before, to the best of the authors' knowledge. Once the room axes have been estimated, identifying first-order image sources becomes relatively straightforward, namely, they are the closest ones to the true source along each of the 6 oriented room axes.

Complementarily to approaches tackling room geometry estimation, a few approaches are focused on estimating surface absorption coefficients or echo amplitudes from recorded signals given the room geometry or image source positions [8], [10], [32], [33], [44], [45]. The approaches in [10], [32], [33], [44] proceed by discretizing the wave equation in both time and space to solve the corresponding sparse inverse problem. The computational burden of discretizing limits these approaches to either frequencies below 500 Hz [10], [32] or 2D rooms [33], [44]. In contrast, the approach in [45] estimate echo amplitudes blindly given their continuous TOAs via least-square optimization. To tackle the high sensitivity of these techniques to geometrical errors, [8] formulates the problem in the magnitude short-time Fourier domain and robustly solves the corresponding non-linear inverse problem with the help of random sampling consensus.

Finally, a relatively recent class of methods replaces some or all of the previously described steps by making use of virtually-supervised deep learning [42], [48]–[50]. While promising results have been reported, these approaches are currently restricted to the acoustic setups simulated in their training data. Moreover, the ability of various training simulation strategies to generalize to a broad enough range of real measurements is an open question that calls for further investigation.

We close this section by observing that most of the above-referenced methods are only tested on a restricted set of geometries. For instance, the methods in [9], [11]–[17], [20]–[23], [25], [27], [28], [30], [46] are tested on less than 3 simulated or real room geometries, and the experimental setups in [18], [20], [21], [23], [37]–[40] have in common a favorable positioning of devices, such as a microphone array near the room center, or sources near the reflectors of interest. Combined with the general absence of publicly available code, this makes existing techniques difficult to reproduce or compare. With the hope of offering a strong baseline, we open-source here an algorithm that is applicable to RIRs simulated by the image source method under fully randomized input parameters, without requiring any hyper-parameter tuning.

III. Image Source Model and Localization

A. The Image Source Method

The ISM [51] relies on a heuristic to approximate the wave equation with impedance boundary conditions. The partial differential equation with boundary conditions is replaced by the following free-field equation containing an image-source *point cloud* as a source term:

$$\frac{1}{c^2}\frac{\partial^2 p}{\partial t^2}(\boldsymbol{r},t) - \Delta p(\boldsymbol{r},t) = \sum_{k=0}^{+\infty} a_k \delta_{\boldsymbol{r}_k}(\boldsymbol{r})\,\delta_0(t). \qquad (1)$$

Each $\boldsymbol{r}_k$ corresponds to the location of a point source emitting an impulse at $t=0$, casting an outward spherical wave. Each of these image-sources is equivalent to a reflection path of the original sound wave on the walls[1]. First order image-sources are constructed by taking the symmetry of the source with respect to the walls, and higher order reflections are obtained by iterating this process.

In the shoebox case, the image-source coordinates are most easily expressed in a *reference frame* of the room, meaning

[1]In the remainder of the manuscript, for convenience, the term *wall* is used to refer to any of the 6 room boundaries, including the floor and the ceiling

a frame composed of an orthonormal basis $(\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3)$ of normal vectors to the walls along with an origin located at one of the room's corners. In such a frame, the image-source coordinates are given by:

$$\{\boldsymbol{r}_{\boldsymbol{q},\boldsymbol{\varepsilon}} = \boldsymbol{\varepsilon} \odot \boldsymbol{v}_{d^{\text{src}}} + 2\boldsymbol{q} \odot \boldsymbol{v}_L \mid \boldsymbol{\varepsilon} \in \{-1; 1\}^3,\ \boldsymbol{q} \in \mathbb{Z}^3\} \quad (2)$$

where $\boldsymbol{v}_L = [L_x, L_y, L_z]^\top$ is the room size vector and $\boldsymbol{v}_{d^{\text{src}}} = [d_x^{\text{src}}, d_y^{\text{src}}, d_z^{\text{src}}]^\top$ gives the distance of the source to each wall containing the origin. Hence, the image sources lie on eight distinct translated orthogonal lattices of common mesh size $2L_x \times 2L_y \times 2L_z$.

Each image source is weighted with an amplitude $a_k \in [0, 1]$ that models the multiplicative decay caused by the reflections of the original sound wave on the walls. The amplitudes of first-order sources correspond to the reflection coefficients of each wall. Note that, as proven in an appendix of the original paper of Allen and Berkley [51], equation (1) is only equivalent to the original wave equation with boundary conditions if each reflection coefficient is equal to one, which corresponds to the case of perfectly reflecting (rigid) boundaries, *i.e*, homogeneous Neumann boundary conditions.

The analytical solution to equation (1) is given by a linear combination of delayed Green functions:

$$p(\boldsymbol{r}, t) = \sum_{k=0}^{+\infty} a_k \frac{\delta\left(t - \|\boldsymbol{r} - \boldsymbol{r}_k\|_2 / c\right)}{4\pi \|\boldsymbol{r} - \boldsymbol{r}_k\|_2}. \quad (3)$$

In practice, p is only observed at some microphone locations $\{\boldsymbol{r}_m^{\text{mic}}\}_{m=1}^M$. Moreover, the observed signal is filtered by the microphones and sampled in time, yielding a discrete measurement vector $\boldsymbol{x} \in \mathbb{R}^{MN}$, defined componentwise by:

$$x_{m,n} \coloneqq \left(\kappa_m * p(\boldsymbol{r}_m^{\text{mic}}, \cdot)\right)(n/f_s) \quad (4)$$

$$= \sum_{k=0}^{K} a_k \frac{\kappa_m\left(n/f_s - \left\|\boldsymbol{r}_m^{\text{mic}} - \boldsymbol{r}_k\right\|_2 / c\right)}{4\pi \left\|\boldsymbol{r}_m^{\text{mic}} - \boldsymbol{r}_k\right\|_2} \quad (5)$$

where f_s is the sampling frequency and $\{\kappa_m\}_{m=1}^M$ are source-microphone response filters. Most ISM implementations use ideal low-pass filters $\kappa_m(t) = \kappa(t) = \operatorname{sinc}(\pi f_s t)$, which will also be the case throughout this article. Each sub-vector $\boldsymbol{x}_m \in \mathbb{R}^N$ corresponds to a discrete and filtered RIR presenting filtered spikes at the times of arrival of image sources. Note that we only consider a finite number of image sources in equation (5). This approximation is reasonable because in practice synthesized RIRs are finite in time and the effect of reflections arriving at microphones later than the final time N/f_s is negligible.

B. Source recovery

In order to estimate image source locations from $\boldsymbol{x}$ we apply the algorithm previously proposed in [47], which is briefly reviewed below. The discrete multichannel RIR $\boldsymbol{x}$ can be expressed as the forward pass of the right hand side of equation (1) through a linear observation operator Γ, *i.e.*

$$\boldsymbol{x} = \Gamma(\sum_{k=0}^{K} a_k \delta_{\boldsymbol{r}_k}). \quad (6)$$

Let $\mathcal{M}_*(\mathbb{R}^3) \subset \mathcal{M}(\mathbb{R}^3)$ be the subset of *Radon measures*[2] that can be written as a linear combination of Dirac masses. The following optimization problem is considered to jointly recover the image source locations and amplitudes:

$$\underset{\psi \in \mathcal{M}_*(\mathbb{R}^3)}{\operatorname{argmin}} \ \|\boldsymbol{x} - \Gamma\psi\|_2^2. \quad (7)$$

This problem is non-convex, but can be relaxed to an infinite-dimensional convex problem by extending it to the whole set of Radon measure $\mathcal{M}(\mathbb{R}^3)$ and by regularizing with a total variation norm on ψ. This relaxed problem can then be addressed by the *sliding Frank-Wolfe* algorithm, a greedy approach proposed by [52], that belongs to the broader class of so-called *super-resolution* or *gridless* techniques. We adapted this method to tackle problem (7) in [47], and were able to accurately recover hundreds of image sources within range, for large enough sampling frequencies and array sizes. Importantly, the recovered image sources are *unlabeled*, some may be missing, and false positives or mislocated sources may exist. Recovering the room parameters from such a noisy, unlabelled image source point cloud is the main contribution of this article, as presented in the following section.

IV. Recovery of Room Parameters

This section presents the proposed room parameter estimation algorithm, consisting in three steps that are detailed in each of the following subsections. The first step is to estimate the room's orientation, the second step is to label the original and first-order image sources, and the third step consists in inferring the remaining parameters, *i.e.*, the source position, the distance of the source relative to each wall (room translation), the room dimensions, and the wall absorption coefficients.

A. Room Orientation

Let us first consider the task of recovering the room orientation from an unlabelled image source point cloud. The key idea is to estimate its underlying orthogonal grid structure, which is apparent in the examples of Fig.1**(a)** and 2. The task amounts to finding a rotation matrix that transforms the microphone array's reference frame to the room's reference frame, up to a permutation of directions. By Eq. (2), the projected coordinates of image sources onto a normal vector to a wall will form clusters, each cluster containing the coordinates of a plane of image sources parallel to this wall. Conversely, projecting image sources onto a randomly chosen vector will, intuitively, not form clusters but instead spread out over the entire range of possible values. In other words, the room basis vectors are orthogonal to the image-source planes generated by the corresponding walls and are expected to maximize the number of orthogonalities. Our method seeks to exploit this structure by scoring basis vector candidates according to their orthogonality to the directions generated by image-source pairs. Formally, let us define f_D as follows:

$$\forall \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^D, \quad f_D(\boldsymbol{u}, \boldsymbol{v}) = \begin{cases} 1 & \text{if } \boldsymbol{u} \perp \boldsymbol{v} \\ 0 & \text{otherwise.} \end{cases} \quad (8)$$

[2] $\mathcal{M}(\mathbb{R}^3)$ is the topological dual of the space of continuous functions on $\mathbb{R}^3$ that vanish at infinity, see [52].

Algorithm 1 Orientation estimation

Input: Image sources $(\boldsymbol{r}_k)_{k=1}^K$
Output: Estimated room orthonormal basis $\hat{\boldsymbol{e}}_1, \hat{\boldsymbol{e}}_2, \hat{\boldsymbol{e}}_3$

$\hat{\boldsymbol{e}}_1 \leftarrow \operatorname{argmin}_{\boldsymbol{u}\in\mathcal{S}^2_{\text{discr}}} J_3^{0.01}$
for $\sigma \in [0.01, 0.005, 0.0005]$ **do**
  $\hat{\boldsymbol{e}}_1 \leftarrow \text{local\_descent}(\hat{\boldsymbol{e}}_1, J_3^\sigma)$
end for
$\hat{\boldsymbol{e}}_2 \leftarrow \operatorname{argmin}_{\boldsymbol{u}\in\mathcal{S}^1_{\text{discr}}} J_{2,\hat{\boldsymbol{e}}_1}^{0.01}$
for $\sigma \in [0.01, 0.005, 0.0005]$ **do**
  $\hat{\boldsymbol{e}}_2 \leftarrow \text{local\_descent}(\hat{\boldsymbol{e}}_2, J_{2,\hat{\boldsymbol{e}}_1}^\sigma)$
end for
$\hat{\boldsymbol{e}}_3 = \hat{\boldsymbol{e}}_1 \times \hat{\boldsymbol{e}}_2$

Let $\mathcal{G} \subset \mathbb{R}^3$ be a finite set of image source locations. Let us consider the following optimization problem:

$$\max_{\|\boldsymbol{u}\|_2=1} J_3(\boldsymbol{u}), \quad \text{where} \quad J_3(\boldsymbol{u}) = \sum_{\boldsymbol{s},\boldsymbol{p}\in\mathcal{G}} f_3(\boldsymbol{u}, \boldsymbol{s}-\boldsymbol{p}). \quad (9)$$

It can be shown that in the noiseless case, for a complete finite cuboid grid $\mathcal{G}$ of image sources, the solution to this problem is indeed a wall normal:

Proposition 1. *Let N_1, N_2, N_3 be non-zero even integers. Consider the following subset of image sources: $\mathcal{G} = \{\boldsymbol{r}_{\boldsymbol{q},\boldsymbol{\varepsilon}},\ \boldsymbol{q} \in [\![0, N_1/2-1]\!]\times[\![0, N_2/2-1]\!]\times[\![0, N_3/2-1]\!],\ \varepsilon_i \in \{-1, 1\}\}$ with $\boldsymbol{r}_{\boldsymbol{q},\boldsymbol{\varepsilon}}$ defined as in* (2)*. Then, a solution $\boldsymbol{u}^*$ to problem (9) is a wall normal,* i.e, *$\boldsymbol{u}^* = \pm\boldsymbol{e}_i$ for some $i \in [\![1, 3]\!]$.*
Proof: See Appendix A.

Note that in Proposition 1, the coordinates are expressed using Eq. (2), *i.e.*, in the unknown reference frame of the room. However, the definition of the cost function J_3 is independent of the coordinate system, so that the result remains true in any coordinate frame. Note also that adversarial cases could be built by carefully removing sources from the image-source point cloud in order to have the score function bear its maximum in a wrong direction. However, assuming the reconstruction algorithm of Section III-B misses image sources at random, the probability of encountering such an adversarial situation is vanishingly small, and Proposition 1 is expected to hold for generic subsets, as will be confirmed by our experiments.

In practice, the image-source reconstruction is noisy and the function f_D defined in (8) cannot be computed exactly. f_D is instead approximated using a Gaussian kernel

$$f_D^\sigma(\boldsymbol{u}, \boldsymbol{v}) = \exp\left(-\frac{1}{2\sigma^2}\left(\frac{\boldsymbol{u}.\boldsymbol{v}}{\|\boldsymbol{u}\|_2\,\|\boldsymbol{v}\|_2}\right)^2\right), \quad (10)$$

such that $\lim_{\sigma\to 0} f_D^\sigma = f_D$ in the pointwise sense. The scale parameter σ controls the tightness of the approximation and plays a regularizing role with respect to the error committed in the localization of image sources. A small σ will yield a noisy loss function if the source localization error is high. Conversely, a large σ means poor precision on room orientation recovery. As we are searching for an optimal *unit* vector,

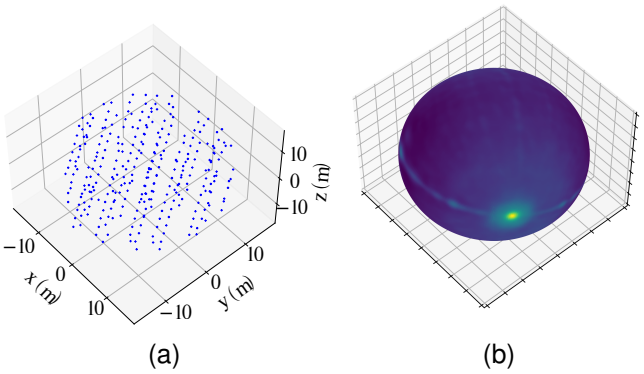

Fig. 1. (a) Reconstructed image-source point cloud using [47] (b) Associated J_3^σ score plotted on the sphere (brighter is higher). A sharp peak is observed in the direction of a wall normal.

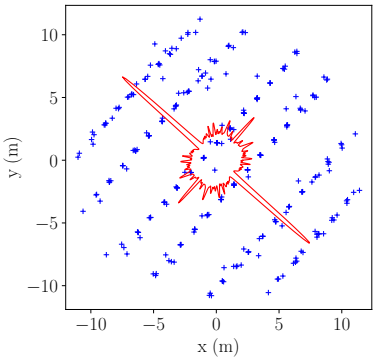

Fig. 2. Projection of the estimated sources on $\hat{\boldsymbol{e}}_1$ (blue) and the associated 2D $J_{2,\hat{\boldsymbol{e}}_1}^\sigma$ score (red). We observe maximal values in the directions of the wall normals.

the regularized score function J_3^σ can be re-parameterized in spherical coordinates by two angles $(\theta, \phi) \in [0, 2\pi[\times[0, \pi[$:

$$J_3^\sigma(\theta, \phi) = \sum_{\boldsymbol{s},\boldsymbol{p}\in\mathcal{G}} f_3^\sigma(\boldsymbol{u}(\theta, \phi), \boldsymbol{s}-\boldsymbol{p}) \quad (11)$$

where $\boldsymbol{u}(\theta, \phi)$ is the unit vector defined by spherical coordinates (θ, ϕ). Once a first basis vector $\boldsymbol{u}$ maximizing J_3^σ has been found, we can proceed in a greedy manner by projecting $\mathcal{G}$ onto $\boldsymbol{u}^\perp$:

$$J_{2,\boldsymbol{u}}^\sigma(\theta) = \sum_{\boldsymbol{s},\boldsymbol{p}\in\mathcal{G}} f_2^\sigma(\boldsymbol{v}(\theta), \mathcal{P}_{\boldsymbol{u}^\perp}(\boldsymbol{s}-\boldsymbol{p})) \quad \forall\theta \in [0, 2\pi[. \quad (12)$$

As can be seen in the examples of Fig. 1 and 2, both score functions J_3^σ and $J_{2,\boldsymbol{u}}^\sigma$ feature maxima along the room axes.

We use the *Scipy* implementation of the *BFGS* algorithm [54] to maximize J_3^σ. Due to the non-convexity of the problem we initialize the optimization algorithm on a finely meshed half-sphere $\mathcal{S}^2_{\text{discr}}$. In order to reduce even more the chance of the algorithm stopping at a local minimum, we begin with a high value of the scale parameter σ and perform the optimization with gradually decreasing values. This process yields an accurate, gridless reconstruction of a first basis vector $\hat{\boldsymbol{e}}_1$, given a sufficiently accurate image-source reconstruction. The sources are then projected onto the plane orthogonal to $\hat{\boldsymbol{e}}_1$ and the process is repeated to recover a second vector $\hat{\boldsymbol{e}}_2$ by

Algorithm 2 Source-wall distances, first order amplitudes

Input: Image sources and amplitudes $(\boldsymbol{r}_k)_{k=1}^K, (a_k)_{k=1}^K$; directions $\hat{\boldsymbol{e}}_1, \hat{\boldsymbol{e}}_2, \hat{\boldsymbol{e}}_3$; threshold μ

Output: Corrected amplitudes up to order 1 $\hat{a}_0, \dots, \hat{a}_6$, source-walls distances $\hat{d}_1, \dots, \hat{d}_6$

$\hat{\boldsymbol{r}}_0 \leftarrow \text{fusion}(\boldsymbol{r}_{k_0}, (a_k)_k, (\boldsymbol{r}_k)_k, \mu),\ k_0 = \text{argmin}_k \|\boldsymbol{r}_k\|_2$
for $t = 1, \dots, 3$ **do**
  $(\boldsymbol{r}_{\text{left}}, \boldsymbol{r}_{\text{right}}) \leftarrow \text{closest\_in\_cone}(\hat{\boldsymbol{r}}_0, \hat{\boldsymbol{e}}_t, (\boldsymbol{r}_k)_k)$
  $(\hat{a}_{t-}, \hat{\boldsymbol{r}}_{t-}) \leftarrow \text{fusion}(\boldsymbol{r}_{\text{left}}, (a_k)_k, (\boldsymbol{r}_k)_k, \mu)$
  $(\hat{a}_{t+}, \hat{\boldsymbol{r}}_{t+}) \leftarrow \text{fusion}(\boldsymbol{r}_{\text{right}}, (a_k)_k, (\boldsymbol{r}_k)_k, \mu)$
end for

optimizing $J^{\sigma}_{2,\hat{\boldsymbol{e}}_1}$. The third vector is then obtained by taking the cross product $\hat{\boldsymbol{e}}_1 \times \hat{\boldsymbol{e}}_2$. The full process is summarized in Algorithm 1. The same values of σ will be used in all of the experiments in this article, without any specific tuning.

B. First order identification and geometry inference

Once the room orientation has been estimated, we seek to identify which of the estimated image sources are of first order. We leverage the fact that the zero-th order image source, *i.e.*, the true source, can be straightforwardly identified. Indeed, it is necessarily the closest one to the microphone array's center. It is also accurately localized, since the direct path is generally well separated from reflections in RIRs. We then cast a cone from the true source in each reconstructed direction $\hat{\boldsymbol{e}}_d$ and their opposite $-\hat{\boldsymbol{e}}_d$. The image source closest to the true source within each cone is picked as a first-order candidate. If the cone is empty (implying that source localization errors are too great) we progressively extend the cone's width until it contains at least one source. As the reconstruction algorithm sometimes produces clusters of sources around the true image-source locations, we assume that any source close to an estimated first order source is a reconstruction artifact. We thus proceed to merge the closest estimated sources. Let $\boldsymbol{r}^*$ be a candidate first-order source, $\mu \in \mathbb{R}_+^*$ a threshold, and $\{\boldsymbol{r}_1^*, \dots \boldsymbol{r}_P^*\}$ the set of reconstructed sources such that $\left\|\boldsymbol{r}_p^* - \boldsymbol{r}^*\right\|_2 < \mu\ \forall p \in [\![1, P]\!]$. We use a heuristic inspired by [55] to merge the corresponding Diracs and their amplitudes:

$$\hat{a} = \sum_{p=1}^{P} a_p^*, \quad \hat{\boldsymbol{r}} = \sum_{p=1}^{P} \frac{a_p^*}{\hat{a}} \boldsymbol{r}_p^*. \tag{13}$$

This procedure gives us estimates for the six first-order image sources and their associated reflection coefficients. The distances of the true source to each wall are then recovered by computing the projections on each estimated wall normal. Let $\hat{\boldsymbol{r}}_{t-}, \hat{\boldsymbol{r}}_{t+}$ be the first order image sources corresponding to $\hat{\boldsymbol{e}}_t$ such that $\hat{\boldsymbol{r}}_{t+}$ is in the cone emitted from $\hat{\boldsymbol{r}}_0$ with direction $\hat{\boldsymbol{e}}_t$ and $\hat{\boldsymbol{r}}_{t-}$ is contained in the opposite cone. The room length in that direction is given by the following formula:

$$\hat{L}_t = \hat{\boldsymbol{e}}_t.(\hat{\boldsymbol{r}}_{t+} - \hat{\boldsymbol{r}}_{t-})/2. \tag{14}$$

Setting the intersection of the walls corresponding to $\hat{\boldsymbol{r}}_{1-}, \hat{\boldsymbol{r}}_{2-}, \hat{\boldsymbol{r}}_{3-}$ as a reference vertex of the room, the translation vector of the room with respect to the source is:

$$\hat{\boldsymbol{\tau}}_{\text{room}} = \frac{1}{2} \begin{pmatrix} \hat{\boldsymbol{e}}_1.(\hat{\boldsymbol{r}}_0 - \hat{\boldsymbol{r}}_{1-}) \\ \hat{\boldsymbol{e}}_2.(\hat{\boldsymbol{r}}_0 - \hat{\boldsymbol{r}}_{2-}) \\ \hat{\boldsymbol{e}}_3.(\hat{\boldsymbol{r}}_0 - \hat{\boldsymbol{r}}_{3-}) \end{pmatrix}. \tag{15}$$

Given the coordinates $\boldsymbol{r}$ of a point in the frame of the microphones, we can then compute the corresponding coordinates in the recovered room frame:

$$\hat{\boldsymbol{r}}_{\text{room}} = \begin{pmatrix} \hat{\boldsymbol{e}}_1^T \\ \hat{\boldsymbol{e}}_2^T \\ \hat{\boldsymbol{e}}_3^T \end{pmatrix} (\boldsymbol{r} - \hat{\boldsymbol{r}}_0) + \hat{\boldsymbol{\tau}}_{\text{room}}. \tag{16}$$

We now have recovered all 18 input parameters that were used to generate the multichannel RIR:

- the room orientation $\hat{\boldsymbol{e}}_1, \hat{\boldsymbol{e}}_2, \hat{\boldsymbol{e}}_3$,
- the 3D source position $\hat{\boldsymbol{r}}_0$,
- the room translation with respect to the source $\hat{\boldsymbol{\tau}}_{\text{room}}$,
- the room dimensions $\hat{L}_1, \hat{L}_2, \hat{L}_3$,
- the 6 wall absorption coefficients $\hat{\alpha}_k = 1 - \hat{a}_k^2$ for $k = 1, \dots, 6$.

Our open-source code for the full image-source reversion procedure is available at: `https://github.com/Sprunckt/acoustic-sfw`.

V. Numerical Experiments

We proceed in this section to evaluate the effectiveness of the proposed inverse algorithm, which can be decomposed into two major steps: first estimating an image source point cloud from a multi-channel RIR and then inferring the room parameters from it. The first step was extensively tested in [47], so we focus here on the estimation of the 18 room parameters given an image-source point cloud estimated using [47], as described in Sec. III-B. All the following tests are based on RIRs simulated using the shoebox ISM, *i.e.*, Eq. (5). As in [47], the RIRs are simulated using image sources up to order 20 and are cut after 50 ms, so that all audible reflections are present in the signals.

A. Simulation Details

We test the full reconstruction procedure on a set of 200 randomly generated rooms containing an omnidirectional impulse sound source and a microphone array. We use two distinct array geometries that are detailed in the following subsections. The rooms' lengths and widths in meters are picked uniformly at random in $[2, 10]$ while the heights are picked in $[2, 5]$. The array's center and the source are randomly placed in each room with a minimal separation distance of one meter to each other and the array is randomly rotated. We also enforce a distance constraint of 25 cm of the array center to each wall to avoid having any microphone placed beyond the room's boundary. Each wall's absorption coefficient is drawn uniformly at random in $[0.01, 0.3]$.

B. Evaluation Metrics

1) Orientation and dimensions: In order to match each recovered direction with the corresponding ground truth wall normal, we apply the ground truth inverse rotation to $(\hat{\boldsymbol{e}}_1, \hat{\boldsymbol{e}}_2, \hat{\boldsymbol{e}}_3)$. Each resulting vector should contain two zero coefficients, the last coefficient being -1 or 1. The indices of the non-zero coefficients allow us to re-order the vectors of the rotation matrix to match the recovered directions. We then compute the mean angular errors between the recovered directions $\hat{e}_d$ and the associated wall normals by taking the arccosine of the dot products. We also compute the mean absolute errors on recovered room dimensions.

2) Wall absorptions: Having matched recovered first-order sources to walls, we compute the mean absolute errors on estimated absorption coefficients: $\hat{\alpha}_k = 1 - \hat{a}_k^2$.

3) Room translation: In order to evaluate the room translation estimation, we calculate the room's center in the array's reference frame from estimated parameters. This is done by inserting $\hat{\boldsymbol{r}}_{\text{room}} = [\hat{L}_1/2, \hat{L}_2/2, \hat{L}_3/2]^\top$ in (16) and solving for $\boldsymbol{r}$. We then calculate the mean of Euclidean distances to the ground truth.

4) RIR extrapolation: Lastly, we evaluate the global accuracy of the method by re-simulating a RIR $\hat{\boldsymbol{x}}$ corresponding to a new random source-array placement in the room using the image-source method (5) with estimated parameters as input. Using the same sampling rate, we compute the signal-to-error ratio to the true RIR $\boldsymbol{x}$ at the new location:

$$\text{SER}(\hat{\boldsymbol{x}}, \boldsymbol{x}) = 10 \log_{10} \left(\frac{\sum_{i=1}^{NM} (\hat{x}_i - x_i)^2}{\sum_{i=1}^{NM} x_i^2} \right). \qquad (17)$$

C. Experimental Results and Analysis

We first consider a 32-element spherical microphone array based on the geometry of the em32 Eigenmike® (radius R=4.2 cm) and scaled by various factors. Figure 3 presents the algorithm's performance on the geometry estimation task for varying sampling frequencies and microphone array radii. In accordance with the image source localization results reported in [47], the accuracy of the estimation improves as the radius or the sampling frequency grow. The lowest resolution ($R = 4.2$ cm and $f_s = 8$ kHz) presents some catastrophic reconstruction failures that heavily impact the mean errors. These catastrophic cases seem to vanish when the array size and sampling rate increase, the mean error steadily converging towards zero for all three metrics. This empirically supports our main claim that the shoebox image-source method is indeed fully algorithmically reversible for large enough arrays and frequencies of sampling. For a frequency of sampling of 24 kHz and the lowest radius, the mean room dimension estimation error is around 3 mm. This number goes down to 0.15 mm when dilating the array by a factor of 5. Meanwhile, the mean error on room orientation (Figure 3.c) remains under 0.06 ° in all experiments, except for the very lowest resolution. The errors on room center localization are somewhat higher. For the smallest array we get a mean error of 0.42 cm at 24 kHz which diminishes to 0.022 cm after dilation. This is expected because estimating the room center couples errors on orientation estimation and source-wall distance estimation.

We then evaluate the estimation of wall absorption coefficients. We observed a few catastrophic absorption recovery even for relatively high array resolutions. In most failure cases, the absorption is grossly overestimated. In order to get a more meaningful picture of the error committed, we only compute the mean errors over coefficients estimated with an error below 0.3 (recall that in our simulations, the coefficients take values in $[0.01, 0.3]$). We also computed the recall rates for this threshold. Both metrics are displayed in Fig. 4. The obtained mean errors are around 0.01, and 100% recall rates are obtained with the largest array sampling at 24 kHz or above. While these are low errors, we do not observe the same convergence towards zero as on geometrical errors. One possible explanation is that we kept the spike estimation algorithm described in [47] untouched, including two spike pruning steps that discard low amplitude Diracs before and after the final gradient descent. While the first pruning step does seem to help the optimization algorithm, the second step, which aimed at reducing false positives, might cause an issue on amplitude estimation. Rather than deleting the spikes and losing the corresponding amplitudes, a lead for improvement would be to merge the spikes by, *e.g*, adapting the heuristic presented in [55].

We now proceed with evaluating the ability of the method to extrapolate RIRs to arbitrary source-array placements in the same room. The results are shown in Fig. 5. Despite the slight absorption errors, we again observe a strong convergence of RIR extrapolation errors towards zero as the array size increases, bringing further support to the claim that the shoebox image-source method has been successfully reversed. Note that we did not observe such convergence as a function of the frequency of sampling. This is expected, since the RIR extrapolation task itself, as assessed by the proposed metric, becomes harder as the frequency of sampling increases. An example of RIR extrapolation result is presented in Fig. 5. As can be seen, the extrapolated RIR very closely matches the ground truth.

We finally study the impact of noise on geometry estimation. Figure 6 presents the recall curves for the recovery of each individual room dimension L_i, for different thresholds. The sampling frequency and array radius are respectively set to 24 kHz and 4.2 cm, and we proceed to varying the peak signal-to-noise ratio (PSNR) of input signals using additive white Gaussian noise. As expected, the algorithm's performance deteriorates when the noise increases and a severe drop appears at 25 dB PSNR. Nevertheless, the algorithm still manages to recover 95.5% of all room dimensions with an error below 5 cm under such noise level, suggesting a reasonable robustness of the overall approach.

D. Baseline Comparison

We now compare the accuracy of the proposed algorithm with the landmark Euclidean distance matrix (EDM)-based method introduced by Dokmanic et al. [22], using the code provided by the authors[3]. This method takes as input a set of unlabelled times of arrival (TOAs) on multiple RIRs, and

[3] https://infoscience.epfl.ch/record/186657/

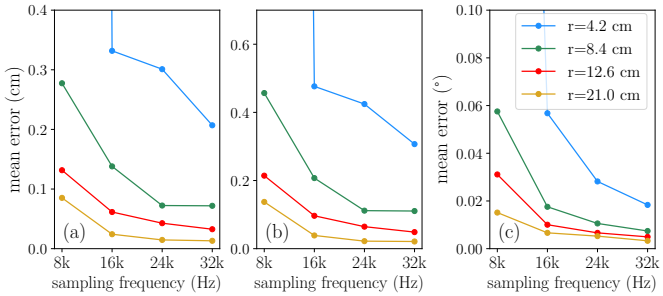

Fig. 3. Mean absolute errors on room dimensions (a), mean Euclidean errors on room center (b) and mean angular error on room orientation (c) in function of the sampling frequency for varying array radii and frequency of sampling.

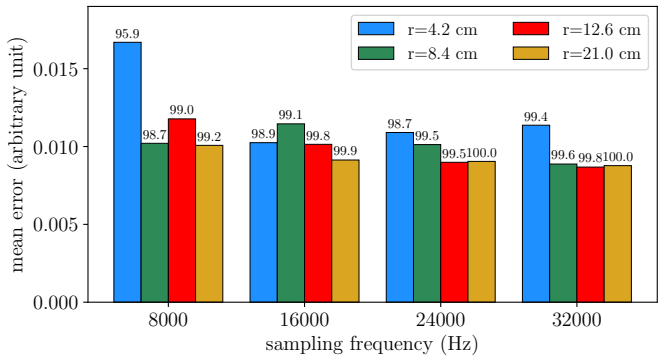

Fig. 4. Mean absolute error on absorption coefficients recovered below a 0.3 threshold for varying array radii and frequency of sampling. The recall for this threshold is indicated above each bar in percent.

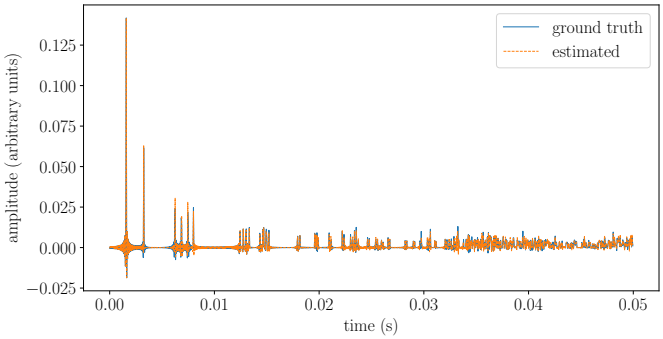

Fig. 5. Example of RIR extrapolation inside the room of Fig. 2 (4.2 cm array radius, 24 kHz frequency of sampling).

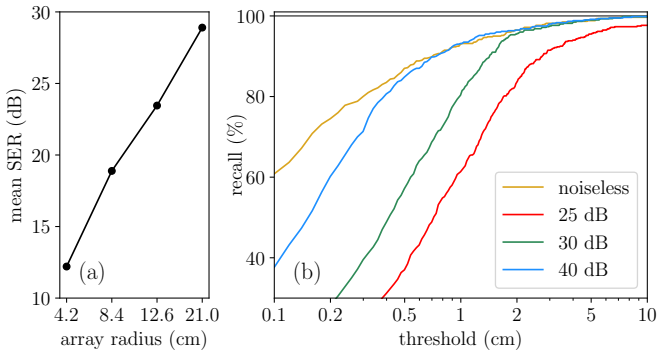

Fig. 6. (a) Mean signal-to-error-ratio of RIR extrapolation for varying array radii at $f_s = 24$ kHz (b) Recall on room dimension recovery as a function of threshold for varying PSNRs for an array radius $r = 4.2$ cm and a frequency of sampling $f_s = 24$ kHz.

returns the 3D locations of first order image sources. Direct comparison on the synthetic dataset defined in Section V-A turned out to be unfeasible. Indeed, the algorithmic complexity of the EDM-based method explodes when the number of reflections increases. Moreover, the method makes the strong assumption that only TOAs from image sources of orders lower than or equal to two are provided. Even when only considering these low-order sources, the number of considered combinations can become very high if the reflections are tightly clustered together due to the room's configuration, which frequently happens in our dataset. Finally, the method was designed for and tested with arrays of typically 5 microphones, since the complexity also drastically increases with the number of channels.

To produce a meaningful comparison, we configure the experiments to be favorable to the EDM-based method. To demonstrate that our approach is agnostic to the array geometry and number of elements, we consider a non-spherical microphone array of 8 microphones composed of two squares stacked on top of each other, the top square being rotated by an angle $\pi/4$. The corresponding array diameter is 37.5 cm. In order to avoid choosing a peak picking method to process the input of the EDM-based method, we place it in an oracle setting. Namely, we provide it with the true times of arrival of all image sources up to order 2 that are in recording range (partial oracle labeling), rounded to the nearest discrete-time sample at 32 kHz. Note that working in discrete time is a fundamental limit of such approaches. We run the two algorithms on the same room configurations as before, only altering the array's geometry but retaining the same location for its center.

TABLE I
RECALL, PRECISION AND MEAN EUCLIDEAN ERRORS (MEE) FOR FIRST-ORDER IMAGE SOURCES (O_1) AND MEE FOR THE TRUE SOURCE (O_0) USING [22] OR THE PROPOSED METHOD.

	O_1 Rec.	O_1 Prec.	O_1 MEE	O_0 MEE
[22]	84.4%	59.7%	65.7± 41.3 mm	35.1± 26.0 mm
Ours	97.2%	97.2%	2.41± 5.71 mm	0.289± 0.584 mm

For each method, we compute the precision and recall for a 20 cm error threshold on the source and first-order image sources localization and labelling. While the proposed algorithm always returns exactly 6 first-order sources, the EDM-based method can wrongfully label second-order reflections as first-order reflections, causing a loss in precision. The results for these experiments are listed in Table I. The localization errors committed by the EDM-based method are an order of magnitude larger than with the proposed approach. This highlights that, even using oracle information, the considered task is far from trivial when considering fully randomized room parameters. The proposed algorithm obtains a mean Euclidean error below 3 mm, which is below $\frac{343}{2\times 32000} \approx 5.1$ mm, the theoretically lowest achievable radial error by any discrete-time

method at this frequency of sampling, indicating that super-resolution is achieved. The number of rooms for which all 6 first-order sources were retrieved without spurious second-order ones was 25.5% for the EDM-based method. Hence, the method could not be used to recover the full geometry of most of the rooms. In contrast, this ratio reached 95.5% of the rooms using the proposed method. For those rooms, the mean geometrical reconstruction errors obtained by it, following the metrics presented in Section V-B, were respectively 0.34 ± 0.6 mm for the room dimensions, 0.61 ± 0.6 mm for the room translation and $0.016 \pm 0.05^\circ$ mm for the room orientation. These are in line with those obtained with the 32-element spherical microphone array of comparable radius and sampling frequency. This seems to indicate that when the array resolution is sufficient, adding microphones does not significantly improve the accuracy of correctly recovered sources. However, adding microphones does seem to reduce some of the geometrical ambiguities and hence to increase the number of correctly identified sources.

VI. Conclusion and Perspective

A new algorithm that leverages the gridless image-source localization method introduced in [47] to achieve full image-source reversion from a discrete, low-passed, multichannel, shoebox RIR was presented. Extensive numerical experiments on simulated RIRs from randomized input parameters reveal that near-exact recovery of all input parameters is achieved by the method, for large enough array sizes and sampling rates. This constitutes, to our knowledge, the first empirical evidence that the historical image-source method of Allen and Berkley [51] is algorithmically reversible, for a wide range of configurations.

The proposed approach is currently not directly applicable to real measured RIRs. This is mainly because the image source localization method it relies upon is specifically designed to reverse the forward image-source model, which makes a number of simplifying assumptions that do not hold in reality. A path towards real-data applicability can nevertheless be envisioned. For this, the method in [47] would need to be extended to take into account both angular and frequency dependencies of receiver, source, and wall responses. Even assuming the responses of the source and microphones are known, and using a physics-based model for the angular dependencies of wall responses, the number of unknown in the problem is then significantly increased. Namely, one needs to additionally estimate the source (and image sources) orientations, as well as a frequency-dependent impedance for each wall. Leveraging additional geometrical and physical constraints on these unknown or incorporating stochastic models are promising leads to make the corresponding inverse problem tractable. Another avenue for future research is to go beyond shoebox geometry. This requires tackling the problem of occlusions, namely, that some image sources may not be visible by all microphones. It also makes the task of room orientation recovery more difficult, and calls for the development of more general point-cloud-to-geometry techniques.

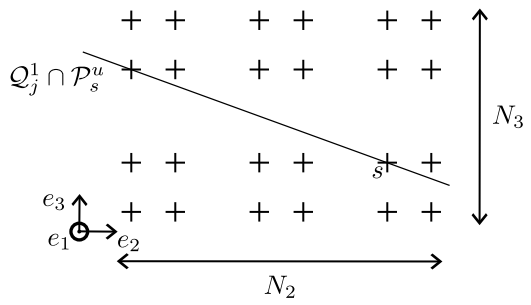

Fig. 7. Intersection of $\mathcal{Q}_j^1$ and $\mathcal{P}_{\boldsymbol{s}}$ when $\boldsymbol{u}$ and $\boldsymbol{e}_1$ are not orthogonal

Appendix A
Proof of Proposition 1

Proof. Note that, by construction, $|\mathcal{G}| = N_1 N_2 N_3$. Let $\boldsymbol{u} \in \mathbb{R}^3$ and denote by $\mathcal{P}_{\boldsymbol{s}}^{\boldsymbol{u}}$ the affine plane passing by $\boldsymbol{s} \in \mathcal{G}$ with normal vector $\boldsymbol{u}$. J_3 can be reinterpreted as the total number of intersections of all planes $\{\mathcal{P}_{\boldsymbol{s}}^{\boldsymbol{u}}\}_{\boldsymbol{s}\in\mathcal{G}}$ with $\mathcal{G}$:

$$J_3(\boldsymbol{u}) = \sum_{\boldsymbol{s}\in\mathcal{G}} |\mathcal{G} \cap \mathcal{P}_{\boldsymbol{s}}^{\boldsymbol{u}}|. \tag{18}$$

Indeed, for all $\boldsymbol{s}, \boldsymbol{p} \in \mathcal{G}$, $\boldsymbol{s}-\boldsymbol{p}$ is orthogonal to $\boldsymbol{u}$ if and only if $\boldsymbol{p} \in \mathcal{P}_{\boldsymbol{s}}^{\boldsymbol{u}}$. Note that for $1 \leq i \leq 3$ the set $\mathcal{G}$ is partitioned by the disjoint union of N_i parallel planes $\mathcal{Q}_j^i \coloneqq \mathcal{P}_{\boldsymbol{s}_j^i}^{\boldsymbol{e}_i}$, $1 \leq j \leq N_i$ where $\{\boldsymbol{s}_j^i, 1 \leq j \leq N_i\} = \{\boldsymbol{r}_{\boldsymbol{q},\boldsymbol{\varepsilon}} \in \mathcal{G}, (q_l, \varepsilon_l) = (0,1) \text{ if } l \neq i\}$. Then:

$$J_3(\boldsymbol{u}) = \sum_{\boldsymbol{s}\in\mathcal{G}} \sum_{j=1}^{N_i} |\mathcal{Q}_j^i \cap \mathcal{P}_{\boldsymbol{s}}^{\boldsymbol{u}} \cap \mathcal{G}| \quad \forall i \in [\![1,3]\!]. \tag{19}$$

Assume in the following that $\boldsymbol{u}$ is not colinear to any of the vectors $\boldsymbol{e}_i$. Then there exists a direction $\boldsymbol{e}_i$ such that every line $\mathcal{Q}_j^i \cap \mathcal{P}_{\boldsymbol{s}}^{\boldsymbol{u}}$, $1 \leq j \leq N_i$ is diagonal, in the sense that the direction of the line is not given by any of the basis vectors $\boldsymbol{e}_j$. Indeed, consider the converse proposition by contradiction: assume that for each $i \in [\![1,3]\!]$ there exists an image source $\boldsymbol{s} \in \mathcal{G}$ and a plane $\mathcal{Q}_j^i$ such that the line $\mathcal{Q}_j^i \cap \mathcal{P}_{\boldsymbol{s}}^{\boldsymbol{u}}$ is generated by a basis vector $\boldsymbol{e}_{k_i}$, $k_i \neq i$. In particular, $\boldsymbol{u}$ is orthogonal to $\boldsymbol{e}_{k_1}$ by definition as $\mathcal{P}_{\boldsymbol{s}}^{\boldsymbol{u}}$ contains the direction $\boldsymbol{e}_{k_1}$. Similarly, $\boldsymbol{e}_{k_{k_1}}$ is orthogonal to $\boldsymbol{u}$. Moreover, as the direction $\boldsymbol{e}_{k_{k_1}}$ is contained in $\mathcal{Q}_j^{k_1}$ which is orthogonal to $\boldsymbol{e}_{k_1}$, then $\boldsymbol{e}_{k_1}$ and $\boldsymbol{e}_{k_{k_1}}$ are distinct. Hence $\boldsymbol{u}$ would be colinear to the last basis vector, raising a contradiction.

We can assume without any loss of generality that direction $\boldsymbol{e}_1$ verifies this property (see Figure 7 for a depiction in that case), *i.e* every line $\mathcal{Q}_j^1 \cap \mathcal{P}_{\boldsymbol{s}}^{\boldsymbol{u}}$, $1 \leq j \leq N_1$ is not generated by $\boldsymbol{e}_2$ or $\boldsymbol{e}_3$. Then, the line $\mathcal{Q}_j^1 \cap \mathcal{P}_{\boldsymbol{s}}^{\boldsymbol{u}}$ intersects $\mathcal{G}$ at at most $\min(N_2, N_3)$ image sources. Moreover, as $\boldsymbol{u}$ is not colinear to $\boldsymbol{e}_2$ and $\boldsymbol{e}_3$, this upper bound can only be reached for the sources $\boldsymbol{s}$ located on the diagonal. Indeed, we need only consider the worst-case scenario, in which the straight line passes through all the nodes on the diagonal. These nodes are at most $\min(N_2, N_3)$. Hence:

$$J_3(\boldsymbol{u}) = \sum_{\boldsymbol{s}\in\mathcal{G}}\sum_{j=1}^{N_1} |\mathcal{Q}_j^1 \cap \mathcal{P}_{\boldsymbol{s}}^{\boldsymbol{u}} \cap \mathcal{G}| < \sum_{\boldsymbol{s}\in\mathcal{G}} N_1 \min(N_2, N_3) = N_1^2 N_2 N_3 \min(N_2, N_3). \quad (20)$$

Now by replacing $\boldsymbol{u}$ with $\boldsymbol{e}_1$, formula (19) becomes :

$$J_3(\boldsymbol{e}_1) = \sum_{j=1}^{N_1} \sum_{\boldsymbol{s}\in\mathcal{G}} |\mathcal{Q}_1^1 \cap \mathcal{G}| \mathbb{1}_{\boldsymbol{s}\in\mathcal{Q}_j^1} = N_1 |\mathcal{Q}_1^1 \cap \mathcal{G}|^2. \quad (21)$$

Thus, $J_3(\boldsymbol{e}_1) = N_1 N_2^2 N_3^2$. We obtain similar formulas for $\boldsymbol{e}_2$ and $\boldsymbol{e}_3$, hence:

$$J_3(\boldsymbol{u}) < \max_{1\leq i\leq 3} J_3(\boldsymbol{e}_i) = N_1 N_2 N_3 \max_{1\leq i<j\leq 3} N_i N_j \quad (22)$$

and the maximum is reached for a vector $\boldsymbol{e}^*$ colinear to $\boldsymbol{e}_1$, $\boldsymbol{e}_2$ or $\boldsymbol{e}_3$. Note that this proof extends to 2D by considering the projection of $\mathcal{G}$ on $\boldsymbol{e}^{*\perp}$ in order to obtain a second basis vector. □

REFERENCES

[1] L. Remaggi, H. Kim, A. Neidhardt, A. Hilton, and P. Jackson, "Perceived quality and spatial impression of room reverberation in vr reproduction from measured images and acoustics," in *Proceedings of ICA*, 2019.

[2] A. Neidhardt, C. Schneiderwind, and F. Klein, "Perceptual matching of room acoustics for auditory augmented reality in small rooms-literature review and theoretical framework," *Trends in Hearing*, vol. 26, p. 23312165221092919, 2022.

[3] A. Canclini, D. Marković, F. Antonacci, A. Sarti, and S. Tubaro, "A room-compensated virtual surround system exploiting early reflections in a reverberant room," in *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*. IEEE, 2012, pp. 1029–1033.

[4] A. Bastine, T. D. Abhayapala, and J. A. Zhang, "Room impulse response reconstruction based on spatio-temporal-spectral features learned from a spherical microphone array measurement," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[5] D. Sundström, F. Elvander, and A. Jakobsson, "Optimal transport based impulse response interpolation in the presence of calibration errors," *IEEE Transactions on Signal Processing*, 2024.

[6] M. Kreković, I. Dokmanić, and M. Vetterli, "Echoslam: Simultaneous localization and mapping with acoustic echoes," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Ieee, 2016, pp. 11–15.

[7] U. Saqib and J. R. Jensen, "A model-based approach to acoustic reflector localization with a robotic platform," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 4499–4504.

[8] S. Dilungana, A. Deleforge, C. Foy, and S. Faisan, "Geometry-informed estimation of surface absorption profiles from room impulse responses," in *30th European Signal Processing Conference (EUSIPCO)*. IEEE, 2022, pp. 867–871.

[9] M. Kuster, D. de Vries, E. Hulsebos, and A. Gisolf, "Acoustic imaging in enclosed spaces: Analysis of room geometry modifications on the impulse response," *The Journal of the Acoustical Society of America*, vol. 116, no. 4, pp. 2126–2137, 2004.

[10] G. P. Nava, Y. Yasuda, Y. Sato, and S. Sakamoto, "On the in situ estimation of surface acoustic impedance in interiors of arbitrary shape by acoustical inverse methods," *Acoustical science and technology*, vol. 30, no. 2, pp. 100–109, 2009.

[11] F. Antonacci, A. Sarti, and S. Tubaro, "Geometric reconstruction of the environment from its response to multiple acoustic emissions," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 2822–2825.

[12] S. Tervo and T. Korhonen, "Estimation of reflective surfaces from continuous signals," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2010, pp. 153–156.

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

[14] A. Canclini, P. Annibale, F. Antonacci, A. Sarti, R. Rabenstein, and S. Tubaro, "From direction of arrival estimates to localization of planar reflectors in a two dimensional geometry," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 2620–2623.

[15] I. Dokmanić, Y. M. Lu, and M. Vetterli, "Can one hear the shape of a room: The 2-d polygonal case," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 321–324.

[16] E. Nastasia, F. Antonacci, A. Sarti, and S. Tubaro, "Localization of planar acoustic reflectors through emission of controlled stimuli," in *2011 19th European Signal Processing Conference*. IEEE, 2011, pp. 156–160.

[17] 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)*. IEEE, 2011, pp. 17–20.

[18] F. Ribeiro, D. Florencio, D. Ba, and C. Zhang, "Geometrically constrained room modeling with compact microphone arrays," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 5, pp. 1449–1460, 2011.

[19] 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 Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 10, pp. 2683–2695, 2012.

[20] H. Sun, E. Mabande, K. Kowalczyk, and W. Kellermann, "Localization of distinct reflections in rooms using spherical microphone array eigenbeam processing," *The Journal of the Acoustical Society of America*, vol. 131, no. 4, pp. 2828–2840, 2012.

[21] J. Filos, A. Canclini, F. Antonacci, A. Sarti, and P. A. Naylor, "Localization of planar acoustic reflectors from the combination of linear estimates," in *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*. IEEE, 2012, pp. 1019–1023.

[22] 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. 12 186–12 191, 2013.

[23] E. Mabande, K. Kowalczyk, H. Sun, and W. Kellermann, "Room geometry inference based on spherical microphone array eigenbeam processing," *The Journal of the Acoustical Society of America*, vol. 134, no. 4, pp. 2773–2789, 2013.

[24] 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.

[25] A. M. Torres, J. J. Lopez, B. Pueo, and M. Cobos, "Room acoustics analysis using circular arrays: An experimental study based on sound field plane-wave decomposition," *The Journal of the Acoustical Society of America*, vol. 133, no. 4, pp. 2146–2156, 2013.

[26] A. H. Moore, M. Brookes, and P. A. Naylor, "Room geometry estimation from a single channel acoustic impulse response," in *21st European Signal Processing Conference (EUSIPCO 2013)*. IEEE, 2013, pp. 1–5.

[27] D. Markovic, F. Antonacci, A. Sarti, and S. Tubaro, "Estimation of room dimensions from a single impulse response," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2013, pp. 1–4.

[28] ——, "Soundfield imaging in the ray space," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 12, pp. 2493–2505, 2013.

[29] L. Zamaninezhad, P. Annibale, and R. Rabenstein, "Localization of environmental reflectors from a single measured transfer function," in *2014 6th International Symposium on Communications, Control and Signal Processing (ISCCSP)*. IEEE, 2014, pp. 157–160.

[30] L. Remaggi, P. J. Jackson, W. Wang, and J. A. Chambers, "A 3d model for room boundary estimation," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 514–518.

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

[32] N. Antonello, T. van Waterschoot, M. Moonen, and P. A. Naylor, "Evaluation of a numerical method for identifying surface acoustic impedances in a reverberant room," in *Proc. of the 10th European Congress and Exposition on Noise Control Engineering*, 2015, pp. 1–6.

[33] N. Bertin, S. Kitić, and R. Gribonval, "Joint estimation of sound source location and boundary impedance with physics-driven cosparse

regularization," in *IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 6340–6344.

[34] M. Crocco and A. Del Bue, "Estimation of TDOA for room reflections by iterative weighted l_1 constraint," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 3201–3205.

[35] I. Jager, R. Heusdens, and N. D. Gaubitch, "Room geometry estimation from acoustic echoes using graph-based echo labeling," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 1–5.

[36] T. Rajapaksha, X. Qiu, E. Cheng, and I. Burnett, "Geometrical room geometry estimation from room impulse responses," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 331–335.

[37] L. Remaggi, P. J. Jackson, P. Coleman, and W. Wang, "Acoustic reflector localization: Novel image source reversion and direct localization methods," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 2, pp. 296–309, 2016.

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

[39] L. Remaggi, H. Kim, P. J. Jackson, F. M. Fazi, and A. Hilton, "Acoustic reflector localization and classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 201–205.

[40] M. Lovedee-Turner and D. Murphy, "Three-dimensional reflector localisation and room geometry estimation using a spherical microphone array," *The Journal of the Acoustical Society of America*, vol. 146, no. 5, pp. 3339–3352, 2019.

[41] D. Di Carlo, C. Elvira, A. Deleforge, N. Bertin, and R. Gribonval, "Blaster: An off-grid method for blind and regularized acoustic echoes retrieval," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 156–160.

[42] W. Yu and W. B. Kleijn, "Room acoustical parameter estimation from room impulse responses using deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 436–447, 2020.

[43] C. Tuna, A. Canclini, F. Borra, P. Götz, F. Antonacci, A. Walther, A. Sarti, and E. A. Habets, "3d room geometry inference using a linear loudspeaker array and a single microphone," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1729–1744, 2020.

[44] Y. Okawa, Y. Watanabe, Y. Ikeda, and Y. Oikawa, "Estimation of acoustic impedances in a room using multiple sound intensities and fdtd method," in *Advances in Acoustics, Noise and Vibration-Proceedings of the 27th International Congress on Sound and Vibration, ICSV*, 2021.

[45] T. Shlomo and B. Rafaely, "Blind amplitude estimation of early room reflections using alternating least squares," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 476–480.

[46] ——, "Blind localization of early room reflections using phase aligned spatial correlation," *IEEE transactions on signal processing*, vol. 69, pp. 1213–1225, 2021.

[47] 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.

[48] 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," in *Forum Acusticum*, 2023.

[49] I. Yeon and J.-W. Choi, "Rgi-net: 3d room geometry inference from room impulse responses in the absence of first-order echoes," *arXiv preprint arXiv:2309.01513*, 2023.

[50] H. N. Bicer, C. Tuna, A. Walther, and E. A. Habets, "Data-driven joint detection and localization of acoustic reflectors," *arXiv preprint arXiv:2402.06246*, 2024.

[51] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[52] Q. Denoyelle, V. Duval, G. Peyré, and E. Soubies, "The Sliding Frank-Wolfe Algorithm and its Application to Super-Resolution Microscopy," *Inverse Problems*, 2019. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01921604

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

[54] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

[55] Y. Traonmilin, J.-F. Aujol, and A. Leclaire, "Projected gradient descent for non-convex sparse spike estimation," *IEEE Signal Processing Letters*, vol. 27, pp. 1110–1114, 2020.