Gridless 3D Recovery of Image Sources from Room Impulse Responses

Tom Sprunck, Antoine Deleforge, Yannick Privat, Cédric Foy

Gridless 3D Recovery of Image Sources from Room Impulse Responses

Tom Sprunck, Antoine Deleforge, Yannick Privat and Cédric Foy

***Abstract*—Given a sound field generated by a sparse distribution of impulse image sources, can the continuous 3D positions and amplitudes of these sources be recovered from discrete, band-limited measurements of the field at a finite set of locations, *e.g.*, a multichannel room impulse response? Borrowing from recent advances in super-resolution imaging, it is shown that this non-linear, non-convex inverse problem can be efficiently relaxed into a convex linear inverse problem over the space of Radon measures in $\mathbb{R}^3$. The new linear operator introduced here stems from the fundamental solution of the wave equation combined with the receivers' responses. An adaptation of the Sliding Frank-Wolfe algorithm is proposed to numerically solve the problem *off-the-grid*, *i.e.*, in continuous 3D space. Idealized simulated experiments show that the approach can recover hundreds of image sources at a rate and accuracy that are not achievable by previous methods, using a compact microphone array and source placed at random in random-sized shoe-box rooms. The impact of noise, sampling rate and array diameter on these results is also examined.**

***Index Terms*—Acoustic reflectors, room shape, sound field, super-resolution, sliding Frank-Wolfe, convex optimization**

I. Introduction

WHEN an omnidirectional point source located at $\boldsymbol{r}_0^{\text{src}} \in \mathbb{R}^3$ emits an impulse $\delta_0(t)$ inside an empty enclosure, the resulting sound pressure field $p : \mathbb{R}^3 \times [0, \infty) \to \mathbb{R}$ obeys the wave equation with a source term together with boundary conditions. While this set of partial differential equations does not admit a general explicit solution, the particular case of a rectangular room with rigid specular boundaries can be treated using the celebrated *image source method* of Allen and Berkley [1], thanks to an equivalence with the following *free field* inhomogeneous wave equation:

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

were c denotes the speed of sound in m/s. Intuitively, the walls of the room have been removed and replaced by an infinite constellation of *image sources* located at $\{\boldsymbol{r}_k^{\text{src}}\}_{k\in\mathbb{N}^*} \subset \mathbb{R}^3$ that synchronously emit the same impulse $\delta_0(t)$ and correspond to iterated spatial reflections of the original source with respect to the walls. While this equivalence only strictly holds for perfectly reflective surfaces and $a_k = 1$, it is commonly generalized by weighing image sources with coefficients $\{a_k\}_{k\in\mathbb{N}^*} \subset [0, 1]$ to account for a proportion of sound energy absorbed by the walls. If the sound field is measured in discrete time by M omnidirectional microphones placed at $\{\boldsymbol{r}_m^{\text{mic}}\}_{m\in[\![1,M]\!]} \subset \mathbb{R}^3$ inside the room, their corresponding sampled signals can be expressed as

$$x_m[n] = (\kappa_m * p(\boldsymbol{r}_m^{\text{mic}}, \cdot))(n/f_s),\ n \in [\![0, N-1]\!] \qquad (2)$$

where $*$ denotes continuous time-domain convolution, f_s is the microphones' frequency of sampling in Hz, n is a discrete time index and $\kappa_m : \mathbb{R} \to \mathbb{R}$ is a time-domain filter modeling the response of microphone m, which may also include the source response. Such signals are called *room impulse responses* (RIRs). Convolving them with any dry source signal can be used to emulate the corresponding reverberant recorded signal. There is an abundant literature on how to measure them in practice [2], [3] and a number of simulators that can compute $\boldsymbol{x} = (x_m[n])_{m,n} \in \mathbb{R}^{MN}$ efficiently[1] given the room dimensions, the wall reflection coefficients and the source and microphone positions, *e.g.* [4]. While this *forward* physical process is very well understood, fully *reversing* it to recover image source parameters $\{\boldsymbol{r}_k^{\text{src}}, a_k\}_{k\in[\![0,K]\!]}$ from $\boldsymbol{x}$ remains an open and active research topic, whose application domains span sound scene navigation [5], auditory augmented reality [6], room acoustic diagnosis [7], and hearing aids [8].

II. Related Work and Contribution

The problem of recovering image sources from measured audio signals can be viewed as a generalization of many tasks that have been independently investigated in the acoustic signal processing literature over the past decade. Estimating the absolute or relative *times of arrival* of image sources at microphones, also known as early *echoes*, is the focus of [9]–[12] and can be of independent interest in the context of *echo-aware* signal processing, as reviewed in [13]. Localizing *reflectors* in the room is equivalent to localizing their corresponding first-order image source together with the true source. Localizing all external reflectors is popularly known as *hearing the shape of the room*. Most studies on this first estimate echoes and/or directions of arrival of image sources, then label and sort them, and finish by triangulation [14]–[22]. Alternatively, [23] proposes a more direct approach based on sparse optimization. Retrieving the coefficients a_k associated to reflectors in frequency bands is studied in [24] and [7], as they relate to their acoustic *impedance*. Finally, recovering image sources within a given range is the focus of recent non-parametric sound-field reconstruction methods [25], [26]. All these tasks can either be performed using RIRs as in [7], [9], [10], [15], [18]–[23], [27] or *blindly* using unknown source signals as in [11], [12], [14]–[17], [24]–[26].

While the above referenced studies developed a rich variety of methodologies, nearly all of them have in common the

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

Tom Sprunck and Antoine Deleforge are with Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France (firstname.name@inria.fr).

Yannick Privat and Tom Sprunck are with IRMA, Université de Strasbourg, CNRS UMR 7501, 67084 Strasbourg, France (yannick.privat@unistra.fr).

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

[1]Note that late reverberation models in cluttered environment should also include scattering and diffraction, but this is not the focus of this article.

definition of a *discrete grid* in 1D time [7], [9], [10], [12], [14]–[22], in 2D space [16], [17], [25]–[27] or in 3D space [23], as well as the use of sparse optimization techniques and/or peak-picking techniques over such grids. This *on-the-grid* paradigm suffers from intrinsic limitations. First, in 3D, the required grid size grows cubically in the desired range and precision. This fundamentally limits the accuracy of current sparse methods under reasonable computational constraints [23], [25], [26]. Second, time-domain peak-picking fails when peaks are overlapping and distorted due to filtering effects such as κ_m. Existing methods address this by using ad-hoc source and microphone placements inside the room [15], [18]–[22]. Third, sparse optimization over a discrete grid fundamentally suffers from the so-called *basis-mismatch* problem [28], [29], requiring the use of ad-hoc post-processing steps.

In parallel, recent theoretical and methodological advances on the general problem of recovering spikes *off-the-grid* have emerged [29]–[35], notably motivated by applications to *super-resolution* in, *e.g.*, fluorescence microscopy [36]. Except for a couple of recent studies on blind echo estimation [11] and anechoic beamforming [37], these advances seem not to have received significant attention from the audio and acoustics communities as of yet. In this article, a connection to this field is established by showing that the non-linear, non-convex inverse problem of recovering $\{\boldsymbol{r}_k^{\text{src}}, a_k\}_{k\in[\![0,K]\!]}$ from $\boldsymbol{x}$ can be relaxed into a convex linear inverse problem over the infinite-dimensional space of Radon measures in $\mathbb{R}^3$. While existing super-resolution applications typically consider Fourier or Laplace operators, a new linear operator derived from the acoustic measurement model (1) and (2) is introduced here. An adaptation of the sliding Frank-Wolfe algorithm [29] to this setting is shown to efficiently solve the problem numerically in continuous 3D space. Under idealized simulated conditions, the method achieves near-exact recovery of hundreds of image sources using one compact microphone array and one source placed at random in random-sized rooms. To the best of the authors' knowledge, this is not achievable by any previously known methodology. Robustness to noise, array size and sampling rate is also examined.

III. Observation Model and Inverse Problem

Equation (1) can be further generalized to an arbitrary *source mass distribution* ψ, yielding

$$\frac{1}{c^2}\frac{\partial^2}{\partial t^2}p(\boldsymbol{r},t) - \Delta p(\boldsymbol{r},t) = \psi(\boldsymbol{r})\delta_0(t), \tag{3}$$

where ψ belongs to the space $\mathcal{M}(\mathbb{R}^3)$ of *Radon measures*, *i.e.*, the topological dual of the space of continuous functions on $\mathbb{R}^3$ that vanish at infinity [29]. Using the linearity of the wave equation, the general solution of (3) is then given by the following spatial convolution product with a Green function:

$$p(\boldsymbol{r},t) = (G(\cdot,t) * \psi)(\boldsymbol{r}) = \int_{\boldsymbol{r}'\in\mathbb{R}^3} \frac{\delta(t-\|\boldsymbol{r}-\boldsymbol{r}'\|_2/c)}{4\pi\|\boldsymbol{r}-\boldsymbol{r}'\|_2}\psi(\boldsymbol{r}')d\boldsymbol{r}'. \tag{4}$$

Intuitively, $G : \mathbb{R}^3 \times \mathbb{R} \to \mathbb{R}$ is a spherical wave centered at the origin propagating outwards at the speed of sound. It corresponds to the fundamental solution of the wave equation, obtained by setting $\psi(\boldsymbol{r}) = \delta_{\boldsymbol{0}}(\boldsymbol{r})$ in (3). Combining (4) and (2), the following expression is obtained for the discrete signal measured by a microphone m placed at $\boldsymbol{r}_m^{\text{mic}}$ observing p:

$$x_m[n] = \int_{\boldsymbol{r}\in\mathbb{R}^3} \gamma_{m,n}(\boldsymbol{r})d\psi(\boldsymbol{r}) = \langle \gamma_{m,n}, \psi\rangle \tag{5}$$

$$\text{where } \gamma_{m,n}(\boldsymbol{r}) \overset{\text{def}}{=} \frac{\kappa\left(n/f_s - \left\|\boldsymbol{r}_m^{\text{mic}} - \boldsymbol{r}\right\|_2/c\right)}{4\pi\left\|\boldsymbol{r}_m^{\text{mic}} - \boldsymbol{r}\right\|_2}. \tag{6}$$

Observe that the *non-linear*, *non-convex* function $\boldsymbol{\gamma} : \mathbb{R}^3 \to \mathbb{R}^{MN}$ can be seen as the representative of an infinite-dimensional *linear* operator[2] $\boldsymbol{\Gamma} : \mathcal{M}(\mathbb{R}^3) \to \mathbb{R}^{MN}$ that maps an arbitrary source mass distribution ψ to its corresponding observation vector $\boldsymbol{x}$. We can now particularize (4) and (5) to a discrete image source distribution $\psi_{\boldsymbol{a},\boldsymbol{r}^{\text{src}}}$ as in (1), yielding:

$$\psi_{\boldsymbol{a},\boldsymbol{r}^{\text{src}}}(\boldsymbol{r}) \overset{\text{def}}{=} \textstyle\sum_{k=0}^K a_k \delta_{\boldsymbol{r}_k^{\text{src}}}(\boldsymbol{r}), \tag{7}$$

$$p(\boldsymbol{r},t) = \textstyle\sum_{k=0}^K a_k \frac{\delta\left(t-\|\boldsymbol{r}-\boldsymbol{r}_k^{\text{src}}\|_2/c\right)}{4\pi\left\|\boldsymbol{r}-\boldsymbol{r}_k^{\text{src}}\right\|_2}, \tag{8}$$

$$x_m[n] = \textstyle\sum_{k=0}^K a_k \frac{\kappa_m\left(n/f_s-\left\|\boldsymbol{r}_m^{\text{mic}}-\boldsymbol{r}_k^{\text{src}}\right\|_2/c\right)}{4\pi\left\|\boldsymbol{r}_m^{\text{mic}}-\boldsymbol{r}_k^{\text{src}}\right\|_2}, \tag{9}$$

$$\boldsymbol{x} = \textstyle\sum_{k=0}^K a_k \boldsymbol{\gamma}(\boldsymbol{r}_k^{\text{src}}) = \boldsymbol{\Gamma}\psi_{\boldsymbol{a},\boldsymbol{r}^{\text{src}}} \in \mathbb{R}^{MN} \tag{10}$$

where $K \in \mathbb{N}\cup\infty$. Notice that the RIR signals in (9) are weighted sums of delayed filters, which is how most image-source simulators are implemented in practice, *e.g.*, [4]. As the image source order increases, their distances and corresponding times of arrival at microphones increase while their weights decrease. The sound pressure field p and the observations $\boldsymbol{x}$ are hence reasonably approximated by considering a *finite* value for K.

Let $\mathcal{M}_*(\mathbb{R}^3) \subset \mathcal{M}(\mathbb{R}^3)$ denote the subset of sparse Radon measures of the form $\psi_{\boldsymbol{a},\boldsymbol{r}^{\text{src}}}$ with $K \in \mathbb{N}$, $\boldsymbol{a} \in \mathbb{R}_+^{K+1}$ and $\boldsymbol{r}^{\text{src}} \in (\mathbb{R}^3)^{K+1}$, *i.e.* measures that are finite positive combinations of spikes. The inverse problem of recovering the amplitudes and positions of these spikes given noisy observations $\boldsymbol{x}$ can be formulated as follows:

$$\underset{\psi_{\boldsymbol{a},\boldsymbol{r}^{\text{src}}}\in\mathcal{M}_*(\mathbb{R}^3)}{\operatorname{argmin}} \left\|\boldsymbol{x} - \boldsymbol{\Gamma}\psi_{\boldsymbol{a},\boldsymbol{r}^{\text{src}}}\right\|_2^2. \tag{11}$$

This belongs to a general class of problems where the goal is to recover the continuous locations of a set of spikes given discrete linear observations over their measure [29]–[35]. Rather than solving (11), which is a non-convex optimization problem on the amplitudes and positions of the image sources, we follow the approach in [29] and consider a convex relaxation to the whole space of Radon measures:

$$\underset{\psi\in\mathcal{M}(\mathbb{R}^3)}{\operatorname{argmin}} \frac{1}{2}\left\|\boldsymbol{x} - \boldsymbol{\Gamma}\psi\right\|_2^2 + \lambda\|\psi\|_{\text{TV}} \tag{12}$$

where $\lambda \in \mathbb{R}_+^*$ is a parameter and $\|\psi\|_{\text{TV}}$ denotes the *total variation* norm of the Radon measure ψ. For a sparse measure $\psi_{\boldsymbol{a},\boldsymbol{r}^{\text{src}}} \in \mathcal{M}_*(\mathbb{R}^3)$ we have $\|\psi_{\boldsymbol{a},\boldsymbol{r}^{\text{src}}}\|_{\text{TV}} = \sum_{k=0}^K |a_k| = \|\boldsymbol{a}\|_1$.

[2]Strictly speaking, this operator is not well defined because $\boldsymbol{\gamma}$ is *singular* at each microphone position. In theory, one should change the integration domain in (5) to $\mathbb{R}^3_\varepsilon \overset{\text{def}}{=} \mathbb{R}^3 \setminus \cup_{m=1}^M B(\boldsymbol{r}_m^{\text{mic}}, \varepsilon)$ for a fixed $\varepsilon > 0$, and only consider measures $\psi \in \mathcal{M}(\mathbb{R}^3_\varepsilon)$. In practice, this adjustment is harmless as long as a minimum separation distance ε is assumed between the image sources and the microphones, and is hence ignored here for clarity.

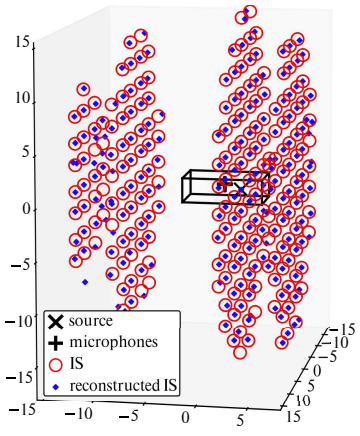

Fig. 1. 3D plot of a room and the corresponding target and reconstructed image sources for a 32-channel spherical microphone array with diameter 16.8 cm, f_s =16 kHz, T_{max} = 50 ms and no noise. The corresponding RIR at one microphone is shown in Fig. 2.

Hence, the second term can be seen as a sparsity-inducing regularizer. By analogy with the finite-dimensional sparse setting, this problem has been coined the Beurling-LASSO (BLASSO) in [30] and offers good measure-reconstruction guarantees in low noise regimes. In particular, [38] shows that there always exists a sparse solution $\psi_{\boldsymbol{a},\boldsymbol{r}^{\text{src}}} \in \mathcal{M}_*(\mathbb{R}^3)$ to problem (12) and [34] studies its basins of attraction.

IV. Algorithm

In order to solve (12) numerically, we adapt the Sliding Frank-Wolfe algorithm proposed in [29], which is briefly reviewed below. Let us denote by $\boldsymbol{a}^{(i)} \in \mathbb{R}_+^{Q_i}$ and $\boldsymbol{r}^{(i)} \in (\mathbb{R}^3)^{Q_i}$ the lists of Q_i spike amplitudes and positions estimated at iteration i, with $\boldsymbol{a}^{(0)} = \boldsymbol{r}^{(0)} = \emptyset$. At iteration $i+1$, the following four steps are performed:

- **Step 1:** A new spike location $\boldsymbol{r}_{Q_i+1}$ is first added to $\boldsymbol{r}^{(i+1)}$ by maximizing the following *dual*, non-convex objective based on the current residual $\boldsymbol{y}^{(i)} \overset{\text{def}}{=} \boldsymbol{x} - \boldsymbol{\Gamma}\psi_{\boldsymbol{a}^{(i)},\boldsymbol{r}^{(i)}}$:

$$\max_{\boldsymbol{r}\in\mathbb{R}^3} \eta^{(i)}(\boldsymbol{r}) \overset{\text{def}}{=} \left[\boldsymbol{\Gamma}^*(\boldsymbol{y}^{(i)})\right](\boldsymbol{r}) = \sum_{m,n} y_m^{(i)}[n]\gamma_{m,n}(\boldsymbol{r}) \quad (13)$$

  where $\boldsymbol{\Gamma}^*$ denotes the *Hermitian adjoint* of $\boldsymbol{\Gamma}$.
- **Step 2:** The whole list of amplitudes $\boldsymbol{a}^{(i+1)}$ is updated by minimizing (12) over $\boldsymbol{a}$ only. This amounts to a classical non-negative LASSO convex optimization problem for which efficient solvers are available, *e.g.*, in [39].
- **Step 3** (*Sliding*): The value of the cost function in (12) is further decreased by jointly refining all the values in $\boldsymbol{a}^{(i+1)}$ and $\boldsymbol{r}^{(i+1)}$ through non-convex local search.
- **Step 4:** The spikes whose amplitudes are lower than a threshold α_{min} are removed from $\boldsymbol{a}^{(i+1)}$ and $\boldsymbol{r}^{(i+1)}$.

A number of modifications are introduced to improve the optimization and reduce the computational time. As in [29], the non-convex maximization in step 1 is carried out using the BFGS algorithm, and is very sensitive to the choice of an initial guess. We use the Scipy implementation and propose to initialize it by first solving the problem over a discrete spatial grid. Because the mass of $\eta^{(i)}$ is tightly contained around the target image sources, a fine grid is required. Covering the entire 3D search region with such a grid would be intractable (over 2 million points). To restrict this region, a moving average over 3 samples is applied to the squared residual signal of each microphone, and the sample $\hat{n}_m$ maximizing these signals for each m is calculated. At least one image source is expected to be located on the spheres $\{S(\boldsymbol{r}_m^{\text{mic}}, cf_s\hat{n}_m)\}_m$. Hence, uniform grids with a mean angular spacing of 5° are built on the spheres corresponding to the 8 microphones with highest peaks, as well as on their neighboring spheres with radii ±5 cm. The initial guess is then the value maximizing $\eta^{(i)}(\boldsymbol{r})$ over the union of these grids (∼40k points). Note that due to the singularity of $\boldsymbol{\gamma}$ mentioned in the previous section, problem (13) does not in fact admit a solution. However, experiments showed that initializing far enough from the microphone array removed this issue in practice. Iterations are stopped when the amplitude of the last recovered spike is below $\alpha_{\text{min}} = 0.01$. To further improve the optimization and reduce computational time, the algorithm is first ran on an early-cut version of the time signals, where echoes are better separated. The resulting spikes are then used as an initialization to run the algorithm on progressively longer signals, and this is repeated until the desired signal length N is reached. Finally, the sliding step 3 is only performed *on the very last iteration*, as suggested in [35], using the parallel bounded BFGS implementation in [40] to preserve positive amplitudes. Spikes with an amplitude less than 0.1 are deleted before and after sliding to decrease the number of false positives. λ is fixed to $3 \cdot 10^{-5}$ throughout the experiments based on a preliminary manual tuning. Our code for this algorithm is available at `https://github.com/Sprunckt/acoustic-sfw`.

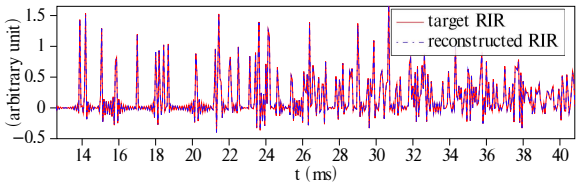

Fig. 2. Excerpt from a room impulse response (RIR) at one microphone and its reconstruction using (9), under the setup described in Fig. 1.

V. Numerical results

We present here some of the numerical results obtained by applying the algorithm described in the previous section to a set of 200 simulated rooms containing an omnidirectional source and a spherical array of 32 microphones. The geometry of the array is the same as the em32 Eigenmike® (diameter d=8.4 cm) but scaled by various factors and using an open-sphere model. We use a unique ideal low-pass filter with cutoff frequency $f_s/2$ to model the response of all the microphones, *i.e.*, $\kappa_m(t) = \text{sinc}(\pi f_s t)$ for all m. The rooms' lengths and widths in meters are sampled uniformly at random in $[2, 10]$, while the heights are taken in $[2, 5]$. The absorption coefficient of each individual wall is sampled uniformly at random in $[0.01, 0.3]$. The source and the array are then placed randomly in each room, with a separation constraint of 1 m to the walls and between each other[3]. While full-lenght RIRs are

[3]This unique constraint was chosen for simplicity, but could be dropped between mic. and walls and safely relaxed between sources and walls.

TABLE I
MEAN ROOM VOLUME ($\overline{\text{V}}$), RECALL (R), PRECISION (P) AND MEAN RADIAL ($\overline{\text{RE}}$), ANGULAR ($\overline{\text{AE}}$), EUCLIDIAN ($\overline{\text{EE}}$) AND AMPLITUDE ($\overline{\text{AmE}}$) ERRORS AMONGST THE RECOVERED SOURCES FOR VARYING NUMBERS OF IMAGE SOURCES, WITH f_s=16 KHZ, d=16.8 CM AND NO NOISE.

# of IS	$\overline{\text{V}}$(m^3)	R(%)	P(%)	$\overline{\text{RE}}$(mm)	$\overline{\text{AE}}$(°)	$\overline{\text{EE}}$(mm)	$\overline{\text{AmE}}$
0-150	214	94.3	81.8	0.069	0.38	94	0.042
150-300	102	92.1	83.1	0.099	0.36	91	0.029
300-500	56	86.1	78.1	0.151	0.38	97	0.025
500-1323	30	57.3	51.6	0.300	0.46	108	0.027

simulated, the proposed approach is only fed with the first $T_{\text{max}} = (N-1)/f_s = 50$ ms of each channel. This allows us to consider as targets all the image sources that are *audible* by all the microphones, *i.e.*, whose distances are inferior to $cT_{\text{max}} = 17.15$ m, independently of the room dimension. For each test room, the ground truth image source positions and amplitudes are obtained using the pyroomacoustics simulator [4] with $c = 343$m/s. An observation vector is then built using (9) and adding white Gaussian noise with a desired peak signal-to-noise ratio (PSNR). An example of room, image source constellation, RIR, recovered image sources and reconstructed RIR is shown in Fig. 1 and Fig. 2.

To evaluate the efficiency of the method, a source is considered *recovered* if at least one estimated source is at an angular distance of less than 2° and a radial distance of less than 1 cm from it with respect to the array center. We then calculate the recall (ratio of true sources recovered), the precision (ratio of estimated sources assigned to a recovered source, discarding doubles), as well as the mean radial, angular and Euclidean errors and the mean error on amplitudes, where the means are calculated *over recovered sources only*. Because the number of image sources that are audible within 50 ms of RIRs varies widely depending on the room's volume, the test set is sliced into four subsets, as detailed in the first two columns of Table I. The remaining columns report the metrics for a sampling frequency f_s=16 kHz, an array diameter d=16.8 cm (x2) and no noise. A recall rate of over 90% for large and medium sized rooms is obtained. The precision is over 80%, indicating few false positives and a reasonable prediction of the number of audible sources. As expected, the recall and precision significantly drop in smaller rooms, where the *echo density* [41] is higher, making the image sources harder to separate. The strength of the proposed gridless approach is revealed by the mean radial and angular errors, which are below tenths of millimeters and fractions of degrees. As a first comparison, the best previously reported results we are aware of in a similar simulated setting are in [24], where an average of 25 nearest image sources are localized with a mean angular error of 4.3°. Note however that [24] is a *blind* method. As a second comparison, ignoring any basis-mismatch issue and assuming *perfect localization*, a sparse method in discrete space such as [23] would require a spatial grid of at least 111 million points to achieve errors below 1° and 1 cm over the same range. This is four orders of magnitude larger than the grids used in the proposed approach.

Notice that the obtained mean Euclidean errors are of a few centimeters. This is because they grow with the source distances, as expected due to the compact spherical geometry of the array. The amplitudes of the recovered sources are also accurately estimated, with mean errors around 0.03 (note that amplitudes lie in $[0, 1]$). These errors are slightly larger in large rooms because amplitudes are larger in that case, due to fewer reflections on the walls. Note that only 3 out of 1200 first-order image sources were missed on this test, while all 200 true sources were recovered.

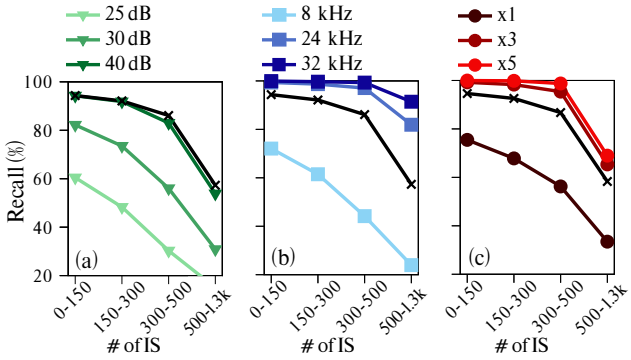

Fig. 3. Recall for varying PSNR (a), sampling frequency (b) and microphone array scaling (c). The default values (×) are noiseless, 16 kHz and x2.

The impact of PSNR, f_s and d on the recall is reported in Fig. 3. Remarkably, it can be observed that either increasing f_s to 32 kHz or the array diameter to 42 cm (x5) brings the recovery rate near 100% for rooms with up to 500 image sources. Conversely, decreasing by half these parameters significantly degrades performance. This is expected as they are known to control the source localization accuracy for compact microphone arrays. Adding noise to the observations does not significantly affect the recovery rate at 40 dB PSNR, but quickly degrades it for PSNRs below 30 dB. Nevertheless, it was observed that the recall values for a PSNR of 30 dB could be restored near the noiseless level by simply considering an angular recovery threshold of 6° instead of 2°. This shows an encouraging stability of the method, given that for such PSNRs the peaks of many echoes in the RIRs fell below the noise standard deviation.

VI. CONCLUSION AND FUTURE WORK

We introduced a new method to recover the continuous 3D positions and amplitudes of all audible image sources given the early part of a discrete-time multichannel RIR from a compact microphone array. While the obtained recovery results under idealized conditions are unprecedented to the best of our knowledge, applying the method to real data will require a number of challenging extensions. Indeed, real RIRs are impacted by the frequency and angular dependencies of source, microphone and wall responses. This will require new formulations of problem (11) in the Fourier and spherical-harmonic domains and an extension of the framework to frequency-dependent amplitudes. Generalization to non-cuboid polyhedral rooms will require robust extensions to *occlusions*, *i.e.*, image sources that are only audible by a subset of microphones. To improve robustness, extending the approach to multiple source and receiver placements in the room is a worthwhile direction. Finally, theoretical investigations on the solutions to problem (12) as well as applications to reflective surface localization and analysis will be pursued.

REFERENCES

[1] Jont B. Allen and David A. Berkley. Image method for efficiently simulating small-room acoustics. *Journal of the Acoustical Society of America*, 65:943–950, 1976.

[2] Guy-Bart Stan, Jean-Jacques Embrechts, and Dominique Archambeau. Comparison of different impulse response measurement techniques. *Journal of the Audio engineering society*, 50(4):249–262, 2002.

[3] Angelo Farina. Advancements in impulse response measurements by sine sweeps. In *Audio engineering society convention 122*. Audio Engineering Society, 2007.

[4] Robin Scheibler, Eric Bezzam, and Ivan Dokmanić. Pyroomacoustics: A python package for audio room simulation and array processing algorithms. In *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 351–355. IEEE, 2018.

[5] Joseph G Tylka and Edgar Choueiri. Comparison of techniques for binaural navigation of higher-order ambisonic soundfields. In *Audio Engineering Society Convention 139*. Audio Engineering Society, 2015.

[6] Annika Neidhardt, Christian Schneiderwind, and Florian Klein. Perceptual matching of room acoustics for auditory augmented reality in small rooms-literature review and theoretical framework. *Trends in Hearing*, 26:23312165221092919, 2022.

[7] Stéphane Dilungana, Antoine Deleforge, Cédric Foy, and Sylvain Faisan. Geometry-informed estimation of surface absorption profiles from room impulse responses. In *30th European Signal Processing Conference (EUSIPCO)*, pages 867–871. IEEE, 2022.

[8] James M Kates. Room reverberation effects in hearing aid feedback cancellation. *The Journal of the Acoustical Society of America*, 109(1):367–378, 2001.

[9] Konrad Kowalczyk, Emanuël AP Habets, Walter Kellermann, and Patrick A Naylor. Blind system identification using sparse learning for TDOA estimation of room reflections. *IEEE Signal Processing Letters*, 20(7):653–656, 2013.

[10] Marco Crocco and Alessio 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)*, pages 3201–3205. IEEE, 2016.

[11] Diego Di Carlo, Clement Elvira, Antoine Deleforge, Nancy Bertin, and Rémi Gribonval. Blaster: An off-grid method for blind and regularized acoustic echoes retrieval. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 156–160. IEEE, 2020.

[12] Tom Shlomo and Boaz Rafaely. Blind localization of early room reflections using phase aligned spatial correlation. *IEEE transactions on signal processing*, 69:1213–1225, 2021.

[13] Diego Di Carlo, Pinchas Tandeitnik, Cedric Foy, Nancy Bertin, Antoine Deleforge, and Sharon Gannot. dEchorate: a calibrated room impulse response dataset for echo-aware signal processing. *EURASIP Journal on Audio, Speech, and Music Processing*, 2021(1):1–15, 2021.

[14] Sakari Tervo and Teemu Korhonen. Estimation of reflective surfaces from continuous signals. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 153–156. IEEE, 2010.

[15] Fabio Antonacci, Jason Filos, Mark RP Thomas, Emanuël AP Habets, Augusto Sarti, Patrick A Naylor, and Stefano Tubaro. Inference of room geometry from acoustic impulse responses. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(10):2683–2695, 2012.

[16] Haohai Sun, Edwin Mabande, Konrad Kowalczyk, and Walter Kellermann. Localization of distinct reflections in rooms using spherical microphone array eigenbeam processing. *The Journal of the Acoustical Society of America*, 131(4):2828–2840, 2012.

[17] Edwin Mabande, Konrad Kowalczyk, Haohai Sun, and Walter Kellermann. Room geometry inference based on spherical microphone array eigenbeam processing. *The Journal of the Acoustical Society of America*, 134(4):2773–2789, 2013.

[18] Ivan Dokmanić, Reza Parhizkar, Andreas Walther, Yue M Lu, and Martin Vetterli. Acoustic echoes reveal room shape. *Proceedings of the National Academy of Sciences*, 110(30):12186–12191, 2013.

[19] Ingmar Jager, Richard Heusdens, and Nikolay D Gaubitch. Room geometry estimation from acoustic echoes using graph-based echo labeling. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2016.

[20] Luca Remaggi, Philip JB Jackson, Philip Coleman, and Wenwu Wang. Acoustic reflector localization: Novel image source reversion and direct localization methods. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(2):296–309, 2016.

[21] Youssef El Baba, Andreas Walther, and Emanuël AP Habets. 3D room geometry inference based on room impulse response stacks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(5):857–872, 2017.

[22] Michael Lovedee-Turner and Damian Murphy. Three-dimensional reflector localisation and room geometry estimation using a spherical microphone array. *The Journal of the Acoustical Society of America*, 146(5):3339–3352, 2019.

[23] Flavio Ribeiro, Dinei Florencio, Demba Ba, and Cha Zhang. Geometrically constrained room modeling with compact microphone arrays. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(5):1449–1460, 2011.

[24] Tom Shlomo and Boaz Rafaely. Blind amplitude estimation of early room reflections using alternating least squares. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 476–480. IEEE, 2021.

[25] Shoichi Koyama and Laurent Daudet. Sparse representation of a spatial sound field in a reverberant environment. *IEEE Journal of Selected Topics in Signal Processing*, 13(1):172–184, 2019.

[26] Stefano Damiano, Federico Borra, Alberto Bernardini, Fabio Antonacci, and Augusto Sarti. Soundfield reconstruction in reverberant rooms based on compressive sensing and image-source models of early reflections. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 366–370. IEEE, 2021.

[27] Luca Remaggi, Hansung Kim, Philip JB Jackson, Filippo Maria Fazi, and Adrian Hilton. Acoustic reflector localization and classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 201–205. IEEE, 2018.

[28] Yuejie Chi, Louis L Scharf, Ali Pezeshki, and A Robert Calderbank. Sensitivity to basis mismatch in compressed sensing. *IEEE Transactions on Signal Processing*, 59(5):2182–2195, 2011.

[29] Quentin Denoyelle, Vincent Duval, Gabriel Peyré, and Emmanuel Soubies. The Sliding Frank-Wolfe Algorithm and its Application to Super-Resolution Microscopy. *Inverse Problems*, 2019.

[30] Yohann De Castro and Fabrice Gamboa. Exact reconstruction using beurling minimal extrapolation. *Journal of Mathematical Analysis and applications*, 395(1):336–354, 2012.

[31] Emmanuel J Candès and Carlos Fernandez-Granda. Towards a mathematical theory of super-resolution. *Communications on pure and applied Mathematics*, 67(6):906–956, 2014.

[32] Vincent Duval and Gabriel Peyré. Exact Support Recovery for Sparse Spikes Deconvolution. *Foundations of Computational Mathematics*, 15(5):1315–1355, 2015.

[33] Veniamin I Morgenshtern and Emmanuel J Candes. Super-resolution of positive sources: The discrete setup. *SIAM Journal on Imaging Sciences*, 9(1):412–444, 2016.

[34] Yann Traonmilin and Jean-François Aujol. The basins of attraction of the global minimizers of the non-convex sparse spike estimation problem. *Inverse Problems*, 36(4):045003, 2020.

[35] Pierre-Jean Bénard, Yann Traonmilin, and Jean-Francois Aujol. Fast off-the-grid sparse recovery with over-parametrized projected gradient descent. In *30th European Signal Processing Conference (EUSIPCO)*, pages 2206–2210. IEEE, 2022.

[36] Bo Huang, Mark Bates, and Xiaowei Zhuang. Super resolution fluorescence microscopy. *Annual review of biochemistry*, 78:993, 2009.

[37] Gilles Chardon and Ulysse Boureau. Gridless three-dimensional compressive beamforming with the sliding frank-wolfe algorithm. *The Journal of the Acoustical Society of America*, 150(4):3139–3148, 2021.

[38] Claire Boyer, Antonin Chambolle, Yohann De Castro, Vincent Duval, Frédéric De Gournay, and Pierre Weiss. On representer theorems and convex regularization. *SIAM Journal on Optimization*, 29(2):1260–1281, 2019.

[39] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[40] Florian Gerber. optimparallel - A parallel version of scipy.optimize. minimize(method='L-BFGS-B'), June 2020. https://doi.org/10.5281/zenodo.3888570.

[41] Helena Peić Tukuljac, Ville Pulkki, Hannes Gamper, Keith Godin, Ivan J Tashev, and Nikunj Raghuvanshi. A sparsity measure for echo density growth in general environments. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2019.