Gridless 3D Recovery of Image Sources from Room Impulse Responses
Tom Sprunck, Antoine Deleforge, Yannick Privat, Cédric Foy

To cite this version:
hal-03763838v2

HAL Id: hal-03763838
https://inria.hal.science/hal-03763838v2
Submitted on 5 Dec 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
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 nonlinear, 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 of-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 $r_0^{\text{src}} \in \mathbb{R}^3$ 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:

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

(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 $\{r_k^{\text{src}}\}_{k \in \mathbb{N}} \subset \mathbb{R}^3$ 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

$$ \{r_m^{\text{mic}}\}_{m \in [1, M]} \subset \mathbb{R}^3 $$

inside the room, their corresponding sampled signals can be expressed as

$$ x_m[n] = (\kappa_m \ast p(r_m^{\text{mic}}, \cdot))(n/f_s), \ n \in [0, N - 1] $$

(2)

where $\ast$ 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 $x = (x_m[n])_{m,n} \in \mathbb{R}^{MN}$ efficiently 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 $\{r_k^{\text{src}}, a_k\}_{k \in [0,K]}$ from $x$ 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 [17], 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]–[22], [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

1Note 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 \(\kappa_m\). 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 \([32]\), 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 \{\[r_{\text{mic}}^k, a_k\] \in [0, K]\} from \(x\) 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 \([4]\) and \([4]\) 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 \(\psi\), yielding

\[
\frac{1}{c^2} \frac{\partial^2}{\partial t^2} p(r,t) - \Delta p(r,t) = \psi(r) \delta_0(t),
\]

where \(\psi\) 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:

\[
p(r,t) = (G(., t) * \psi)(r) = \int_{r' \in \mathbb{R}^3} \frac{\delta(t - ||r - r'||/c)}{4\pi ||r - r'||^2} \psi(r') \, dr'.
\]

Intuitively, \(G : \mathbb{R}^3 \times \mathbb{R} \rightarrow \mathbb{R}\) 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 \(\psi(r) = \delta_0(r)\) in \([3]\). Combining \([4]\) and \([2]\), the following expression is obtained for the discrete signal measured by a microphone \(m\) placed at \(r_{\text{mic}}^m\) observing \(p\):

\[
x_m[n] = \int_{r \in \mathbb{R}^3} \gamma_m(n, r) d\psi(r) = \langle \gamma_m, n, \psi \rangle
\]

where \(\gamma_m(n, r) \equiv \kappa \left( \frac{n/f_s - \|r_{\text{mic}}^m - r\|/c}{4\pi \|r_{\text{mic}}^m - r\|^2} \right)\).

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

\[
\psi_{a, r_{\text{mic}}} (r) = \sum_{k=0}^{K} a_k \delta_{r_{\text{mic}}^k} (r),
\]

\[
p(r, t) = \sum_{k=0}^{K} a_k \delta(t - ||r - r_{\text{mic}}^k||/c),
\]

\[
x_m[n] = \sum_{k=0}^{K} a_k \kappa \left( \frac{n/f_s - \|r_{\text{mic}}^m - r_{\text{mic}}^k\|/c}{4\pi \|r_{\text{mic}}^m - r_{\text{mic}}^k\|^2} \right),
\]

\[
x = \sum_{k=0}^{K} a_k \gamma(r_{\text{mic}}^k) = \Gamma \psi_{a, r_{\text{mic}}} \in \mathbb{R}^{MN}
\]

where \(K \in \mathbb{N} \cup \{\infty\}\). Notice that the RIR signals in \([4]\) 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 \(x\) are hence reasonably approximated by considering a finite value for \(K\).

Let \(\mathcal{M}_x(\mathbb{R}^3) \subseteq \mathcal{M}(\mathbb{R}^3)\) denote the subset of sparse Radon measures of the form \(\psi_{a, r_{\text{mic}}}\) with \(K \in \mathbb{N}, a \in \mathbb{R}_+^{K+1}\) and \(r_{\text{mic}}^k \in (\mathbb{R}^3)^{K+1}\), 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 \(x\) can be formulated as follows:

\[
\arg\min_{\psi_{a, r_{\text{mic}}} \in \mathcal{M}_x(\mathbb{R}^3)} \|x - \Gamma \psi_{a, r_{\text{mic}}} \|^2_2.
\]

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:

\[
\arg\min_{\psi \in \mathcal{M}(\mathbb{R}^3)} \frac{1}{2} \|x - \Gamma \psi\|^2_2 + \lambda \|\psi\|_{TV}
\]

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

\(2\)Strictly speaking, this operator is not well defined because \(\gamma\) is singular at each microphone position. In theory, one should change the integration domain in \([3]\) to \(\mathbb{R}^3 \setminus \cup_{k=0}^{K} B(r_{\text{mic}}^k, \varepsilon)\) for a fixed \(\varepsilon > 0\), and only consider measures \(\psi \in \mathcal{M}(\mathbb{R}^3)\). In practice, this adjustment is harmless as long as a minimum separation distance \(\varepsilon\) is assumed between the image sources and the microphones, and is hence ignored here for clarity.
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 $\psi_{a,r^*} \in \mathcal{M}_s(\mathbb{R}^3)$ 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 $a^{(i)} \in \mathbb{R}^Q$ and $r^{(i)} \in (\mathbb{R}^3)^Q$, the lists of $Q$ spike amplitudes and positions estimated at iteration $i$, with $a^{(0)} = r^{(0)} = \emptyset$. At iteration $i + 1$, the following four steps are performed:

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

  $$\max_{r \in \mathbb{R}^3} \eta^{(i)}(r) \triangleq \frac{1}{2} \left[ \Gamma^{*}(y^{(i)}) \right](r) = \sum_{m,n} \gamma^{(i)}_{m,n}(r) \quad (13)$$

  where $\Gamma^*$ denotes the Hermitian adjoint of $\Gamma$.

- **Step 2:** The whole list of amplitudes $a^{(i+1)}$ is updated by minimizing (12) over $a$ 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 $a^{(i+1)}$ and $r^{(i+1)}$ through non-convex local search.

- **Step 4:** The spikes whose amplitudes are lower than a threshold $a_{\min}$ are removed from $a^{(i+1)}$ and $r^{(i+1)}$.

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 $\bar{\eta}_m$ maximizing these signals for each $m$ is calculated. At least one image source is expected to be located on the spheres $\{S(r^{mic}_m, c f_s \bar{\eta}_m)\}_m$. 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 $\eta^{(i)}(r)$ over the union of these grids (~40k points). Note that due to the singularity of $\gamma$ 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 $a_{\min} = 0.01$. To further improve the optimization and reduce computational time, the algorithm is first run 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. $\lambda$ 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.

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-space model. We use a unique ideal low-pass filter with cutoff frequency $f_c/2$ to model the response of all the microphones, i.e., $\kappa_m(t) = \sin(\pi f_c t)$ 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. While full-length RIRs are

---

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 (V), PRECISION (P) AND MEAN RADIAL (RE), ANGULAR (AE), EUCLIDEAN (EE) AND AMPLITUDE (AME) ERRORS AMONGST THE RECOVERED SOURCES FOR VARYING NUMBERS OF IMAGE SOURCES, WITH f_s=16 kHz, d=16.8 cm AND NO NOISE.

<table>
<thead>
<tr>
<th># of IS</th>
<th>V(m^3)</th>
<th>R(%)</th>
<th>P(%)</th>
<th>RE(mm)</th>
<th>AE(°)</th>
<th>EE(mm)</th>
<th>AME</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-150</td>
<td>214</td>
<td>94.3</td>
<td>81.5</td>
<td>0.069</td>
<td>0.38</td>
<td>94</td>
<td>0.042</td>
</tr>
<tr>
<td>150-300</td>
<td>102</td>
<td>92.1</td>
<td>83.1</td>
<td>0.099</td>
<td>0.36</td>
<td>91</td>
<td>0.029</td>
</tr>
<tr>
<td>300-500</td>
<td>56</td>
<td>86.1</td>
<td>78.1</td>
<td>0.151</td>
<td>0.38</td>
<td>97</td>
<td>0.025</td>
</tr>
<tr>
<td>500-1323</td>
<td>30</td>
<td>57.3</td>
<td>51.6</td>
<td>0.300</td>
<td>0.46</td>
<td>108</td>
<td>0.027</td>
</tr>
</tbody>
</table>

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

simulated, the proposed approach is only fed with the first T_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_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 pyroacoustics simulator [2] 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 [11] 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 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.

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


