# Geometrically Constrained Room Modeling with Compact Microphone Arrays

Flávio Ribeiro*, *Student Member, IEEE,* Dinei Florêncio, *Senior Member, IEEE,* Demba Ba and Cha Zhang, *Senior Member, IEEE*

***Abstract*—The geometry of an acoustic environment can be an important information in many audio signal processing applications. To estimate such a geometry, previous work has relied on large microphone arrays, multiple test sources, moving sources or the assumption of a 2D room. In this paper, we lift these requirements and present a novel method that uses a compact microphone array to estimate a 3D room geometry, delivering effective estimates with low-cost hardware. Our approach first probes the environment with a known test signal emitted by a loudspeaker co-located with the array, from which the room impulse responses (RIRs) are estimated. It then uses an $\ell_1$-regularized least-squares minimization to fit synthetically generated reflections to the RIRs, producing a sparse set of reflections. By enforcing structural constraints derived from the image model, these are classified into 1st, 2nd and 3rd-order reflections, thereby deriving the room geometry. Using this method, we detect walls using off-the-shelf teleconferencing hardware with a typical range resolution of about 1 cm. We present results using simulations and data from real environments.**

***Index Terms*—Room geometry estimation, wall localization, reflector localization, circular microphone array, l1-regularized least squares.**

## I. Introduction

The knowledge of the geometry of an acoustic environment can be helpful in many audio signal processing applications. Indeed, it can be used to increase the accuracy of 3D sound source localization with compact arrays [1], improve localization perception in 3D sound spatialization [2], increase robustness in MVDR beamformers, and initialize acoustic echo cancellation algorithms. In general, one can use the obtained geometric information to analytically determine an arbitrary point-to-point impulse response (using, for example, the image method [3]). The computed impulse responses can be used to compensate early reflections, which would otherwise corrupt a signal of interest.

The problem of extracting the 3D geometry from real-world measurements has been an active area of research for decades, particularly in the fields of machine vision, remote sensing and robotics [4]. However, while vision-based techniques are reasonably effective for extracting detailed geometry information, they tend to create overly complex room models for audio signal processing, since for most audio applications we are only interested in a model of acoustically dominant reflectors. Small or thin objects tend to be acoustically transparent, and can be safely ignored. Furthermore, acoustically reflective materials are not always clearly identifiable through visual inspection.

In this paper, we explore schemes to estimate a room model using acoustic methods. Research along this direction has been attracting interest in recent years. For instance, Moebus and Zoubir [5] performed acoustic imaging with an ultrasound transmitter/receiver pair mounted on a 2D positioning system, by synthesizing a 400-element array and using MVDR beamforming to reveal the position and outline of obstacles. O'Donovan et al. [6] used acoustic imaging with a 32-microphone spherical array to visualize the delay and direction of arrival of sound reflections in concert halls.

Several methods used some variation of the image model [3] and represented the room as a collection of planar reflectors [7]–[11]. One such proposal [7] used a single microphone and a loudspeaker moving over a circular trajectory while emitting white noise. A likelihood map was generated from the cross-correlation of the test signal and the measured response, and used to identify vertical reflectors. In [8], the authors used a loudspeaker moving over a discrete collection of coordinates. Room impulse responses (RIRs) were collected at each location, from which times of arrival (TOAs) were extracted with a peak-picking algorithm. For a fixed reflector and known source position, the locus of all reflection points which produce a given TOA is an ellipse. By using multiple source positions, a reflector can be identified as the common tangent to all ellipses.

Related work used a microphone array and multiple sources to identify a single vertical reflector [9]. For each source location (which must be known), the authors used an MVDR pseudospectrum to estimate the direction of arrival (DOA) of the reflection. The reflector is known to be tangent to a parabola which has the focus on the source and a directrix given by the measured DOA. Using multiple sources, one arrives at a nonlinear least-squares problem, which produces an estimate of the reflector coordinates.

In [10], the authors assumed only one source and one microphone, both stationary, and a 2D environment. They established a matrix relationship between wall normals, 1st-order and 2nd-order image sources. They then used a peak-picking algorithm on the estimated RIRs to find dominant reflections, and searched for the correct permutation of their subset of reflections that satisfies the matrix constraint. Since this subset contains all 1st-order image sources, it directly produces the desired geometry.

A blind two-step estimation method was proposed in [11]. Using a microphone array, the authors first estimated the range and direction of a source using a least-squares fit over measured time differences of arrival. RIRs were then produced by blind estimation, and the location of each wall was inferred from the common tangent approach, in a manner resembling

[8].

In this paper we consider the problem of fitting a room model composed of an arbitrary number of vertical walls, a floor and a ceiling, using an array of $M$ microphones with an integrated loudspeaker. We assume that the array has a fixed and known geometry, and no moving parts. We consider that the array is small enough to be portable and thus produced as a consumer product.

This work is distinguished by three major contributions. The first is the removal of strong assumptions needed by previous methods. An off-the-shelf microphone array typically has a small number of microphones and a compact planar geometry. These arrays produce poor angular resolution, precluding the use of acoustic imaging. Due to the small interelement distances and planar array geometry, methods based on beamforming would have practically no range resolution and would also become unsuitable. Finally, our proposal does not require moving sound sources or microphone arrays around the room, making it easy to implement.

The second contribution is the consideration of a 3D environment with a floor and a ceiling. Most of the image model methods proposed in the literature assume a 2D room, where only walls are considered. While the generalization to 3D space is conceptually straightforward, the floor and ceiling generate a significant number of high order reflections which are non-trivial to address, considering that the order of each reflection (corresponding to a peak in the RIR) is not known a priori.

The third contribution concerns robustness to real-world non-idealities. In practice, white noise is not the dominant source of distortion, as often used in the literature (and in this paper's simulations) to synthesize more challenging scenarios. In real environments, one encounters a multitude of obstacles which are reflective due to their proximity to the test source, but are not acoustically dominant for most possible source locations. Waveforms are also distorted by frequency-dependent reflection coefficients, finite walls, and the coupling of the source with surfaces such as tables. Thus, most of the peaks present in real RIRs do not correspond to dominant reflectors, and the ones that do have been distorted. This creates robustness problems for peak-picking algorithms, even under high SNR scenarios.

To identify strong reflectors, we propose to use an $\ell_1$-regularized least-squares procedure and fit known reflection templates to measured RIRs. The $\ell_1$ regularization promotes sparsity and offers robustness to measurement noise, to device model deviations (such as microphone mismatches and frequency response deviations) and to environment parameters (such as reflector size and surface material). It thus produces a sparse set of strong reflections with known DOA and range. These reflections are analyzed and further classified into 1st, 2nd and 3rd-order reflections or clutter, from which the room model can be correctly inferred.

This paper is organized as follows: Section II gives an overview of the problem and the main assumptions under consideration. Section III presents the mathematical details of the signal model, the impulse response decomposition used to produce wall candidates and the post-processing procedure used to validate them. Section IV describes how to build an array model for synthesizing reflections from arbitrary DOAs, and how to implement a transform to efficiently decompose the measured impulse response into its dominant reflections. Section V shows examples with simulations and real data acquired in corporate environments, featuring surfaces of diverse materials and obstacles such as chairs, cabinets, projectors and light fixtures. Section VI has our conclusions.

## II. Problem statement

We wish to obtain a room model which can be used to predict approximately how sound propagates in a room. The room model need not be perfect, since only the strong early reflections will be accounted for in the applications of interest. Real rooms are potentially complex environments – yet, in sampling a few conference rooms in corporate environments, we find that almost every room has four walls, a ceiling and a floor; the floor is leveled and the ceiling is parallel to the floor; walls are vertical, straight, and extend from floor to ceiling and from adjoining wall to adjoining wall. Carpet is common, and almost invariably there is a conference table in the center of the room. Furthermore, many objects that seem visually important are small enough that may actually be acoustically transparent for most frequencies of interest. Based on these observations, we adopt a simple room model: an arbitrary number of vertical walls, a floor and a ceiling.

Even with such a simplified room model, it would be difficult to blindly estimate the components of the model based solely on unknown signals already existing in the room (as proposed in prior work [11]). Indeed, blind channel estimation is particularly challenging, especially given the very long acoustic impulse responses that are frequently found in practice. Also note that blind channel estimation is inherently ambiguous to group delay. Thus, it requires independent estimates of the direct path length from the source to each of the microphones. This reduces to determining the location of the source, which by itself is a challenging problem due to the presence of multipath and reverberation.

Instead, we follow the same approach as [7], [8], [10] and actively probe the room by emitting a known signal (in our case, a sine sweep) from a source at a known location. We assume that the source and microphones are synchronized. For the purposes of this discussion, we consider a uniform circular microphone array with a speaker rigidly mounted in its center. Nevertheless, any sufficiently diverse geometry suffices. We only assume that the array is small and has a known geometry, allowing us to use a computationally efficient plane-wave propagation model.

Note that in contrast to previous work, we use a single sound source located close to the microphones. This implies that we only sample each wall at the point where its normal vector points to the array. Thus, we assume that the walls extend beyond the location at which they are detected. Fig. 1 illustrates the concept when using the proposed room model for speech enhancement or sound source localization. The circular device in the room detects the reflections from the walls, indicated by the black segments in each of the four

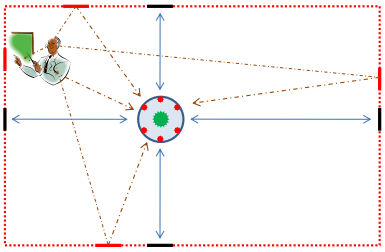


Figure 1: Reflection model.

walls. However, the locations of interest for the walls are in fact the ones indicated by the red segments. The underlying assumption is that the walls extend linearly and with similar acoustic characteristics.

We consider the problem of fitting a model of N planar reflectors to a 3D enclosure, using impulse responses estimated with an M-microphone array. The room model is denoted $\mathcal{M} = \{(r_i, \theta_i, \phi_i)\}_{i=1}^{N}$, where the vector $(r_i, \theta_i, \phi_i)$ specifies respectively the range, azimuth and elevation of the $i^{th}$ reflector with respect to a known coordinate system. We define the coordinate system such that for all side walls, $\phi_i = 0°$. For the ceiling and floor, $\phi_i = 90°$ and $\phi_i = -90°$, respectively.

The obvious model fitting approach would be completely parametric, where $\mathcal{M}$ is estimated directly by minimizing an objective function. However, while one can use the image method to obtain RIRs from $\mathcal{M}$, the map from the RIRs to the parameter space is highly nonlinear and difficult to optimize. Furthermore, real RIRs are certain to have missing reflections due to occlusion, added features due to clutter, and distortions due to deviations from the device and propagation models. These would cause a significant departure from the parametric model, and would likely lead to estimates with major errors. Thus, we resort to a non-parametric method which assumes that early segments of impulse responses can be decomposed into a sum of isolated wall reflections, which can be independently identified and later cross-validated using a post-processing procedure.

## III. Room Modeling

### A. Definitions

In the following, we will use $\wedge$ and $\neg$ to denote the logical conjunction and negation operators.

Without loss of generality, a spherical coordinate system $(r, \theta, \phi)$ is defined such that $r$ is the range, $\theta$ is the azimuth, $\phi$ is the elevation and $(0, 0, 0)$ coincides with the loudspeaker. We assume that the geometry of the array and loudspeaker are fixed and known a priori.

Reflectors are identified by a 3D point $(r, \theta, \phi)$, under the assumption that the reflector aligns with the tangential plane at point $(r, \theta, \phi)$ of the sphere with radius $r$ centered at the origin. Even though a high order reflection involves multiple reflecting surfaces, their combination is equivalent to a single virtual reflector. Thus, high order reflections are represented in the same manner as 1st-order reflections. Note that the 1st-order image source with respect to a reflector $(r, \theta, \phi)$ is located at $(2r, \theta, \phi)$.

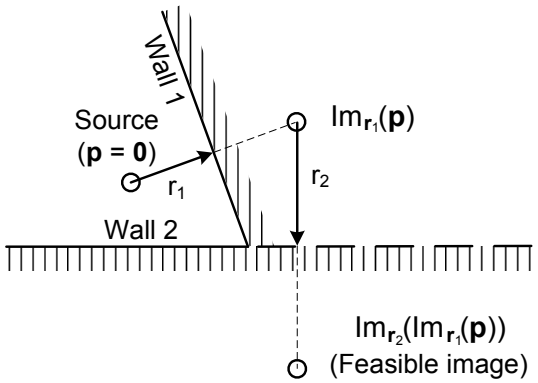



Figure 2: Example of a feasible 2nd-order image.

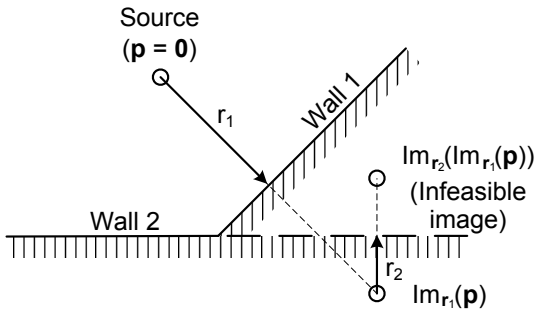



Figure 3: Example of an infeasible 2nd-order image.

Let $\boldsymbol{r} \in \mathbb{R}^3$ be a reflector normal in Cartesian coordinates. Define the image of a source located at $\boldsymbol{p} \in \mathbb{R}^3$ with respect to $\mathbf{r}$ by

$$\mathrm{Im}_{\mathbf{r}}(\mathbf{p}) = \mathbf{p} + 2\left(1 - \frac{\mathbf{r}^T\boldsymbol{p}}{\|\mathbf{r}\|^2}\right)\mathbf{r}. \tag{1}$$

Denote the 1st, 2nd and 3rd-order reflections with respect to reflectors $\mathbf{r}_1, \mathbf{r}_2$ and $\mathbf{r}_3$ by

$$\begin{aligned}
\Xi(\mathbf{r}_1) &= \frac{1}{2}\mathrm{Im}_{\mathbf{r}_1}(\mathbf{0}) \\
\Xi(\mathbf{r}_1, \mathbf{r}_2) &= \frac{1}{2}\mathrm{Im}_{\mathbf{r}_2}(\mathrm{Im}_{\mathbf{r}_1}(\mathbf{0})) \\
\Xi(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3) &= \frac{1}{2}\mathrm{Im}_{\mathbf{r}_3}(\mathrm{Im}_{\mathbf{r}_2}(\mathrm{Im}_{\mathbf{r}_1}(\mathbf{0}))).
\end{aligned}$$

It can be shown that $\left(1 - \frac{\mathbf{r}^T\boldsymbol{p}}{\|\mathbf{r}\|^2}\right)$ in (1) is the signed point-plane distance divided by $\|\mathbf{r}\|$. An image source is infeasible if this signed point-plane distance is negative for any image source in the acoustic path, indicating that a reflection occurred along the non-reflective side of the boundary. While 1st-order reflections are always feasible, higher order reflections should be checked. Fig. 2 shows an example of a feasible 2nd-order image. Fig. 3 shows an example of an infeasible 2nd-order image, where the reflection across Wall 2 happens over the non-reflective side (in fact, all 2nd-order images between reflectors with obtuse angles are infeasible).

A second test for feasibility requires testing whether each reflection occurs within the limits of its corresponding walls [12]. In this paper, we only consider 2nd and 3rd-order reflections involving at most two walls and the ceiling. It is possible to show that such reflections are always feasible if the walls are assumed to be infinite and perpendicular to the

ceiling. Thus, since we do not know a priori the size of each wall, this visibility test does not apply.

For reflections $\mathbf{r}_1 = (r_1, \theta_1, \phi_1), \mathbf{r}_2 = (r_2, \theta_2, \phi_2)$, define the difference function $\Delta$ as

$$\begin{aligned}\Delta(\mathbf{r}_1, \mathbf{r}_2) &= (|r_1 - r_2|, |\theta_1 - \theta_2|_\theta, |\phi_1 - \phi_2|) \\ |\theta_1 - \theta_2|_\theta &= \min(|\theta_1 - \theta_2|, 360^\circ - |\theta_1 - \theta_2|).\end{aligned}$$

We extend the $<$ relation to spherical coordinates such that

$$\mathbf{r}_1 < \mathbf{r}_2 \Leftrightarrow (r_1 < r_2) \wedge (\theta_1 < \theta_2) \wedge (\phi_1 < \phi_2).$$

Let $\boldsymbol{\delta} = (\delta_r, \delta_\theta, \delta_\phi)$ be the array resolution threshold for reflections. Thus, if $\Delta(\mathbf{r}_1, \mathbf{r}_2) < \boldsymbol{\delta}$, then $\mathbf{r}_1$ and $\mathbf{r}_2$ can be considered to be generated by the same image source.

### B. Signal model

Define $h_m^{(r,\theta,\phi)}(n)$ as the discrete time impulse response from the loudspeaker to the $m^{th}$ microphone, considering that: (1) the direct path from the loudspeaker to the microphone has been removed and (2) the array is mounted in free space, except for the presence of a lossless, infinite wall passing through point $(r, \theta, \phi)$ with normal vector $\mathbf{n} = (\theta, \phi)$. Let $r$ be sufficiently large so that the wall does not intersect the array or offer significant near-field effects. We call $h_m^{(r,\theta,\phi)}(n)$ a wall impulse response (WIR), where $n$ is the sample index, and $m$ is the microphone index.

Our discrete time observation model is

$$y_m(n) = h_m(n) * s(n) + u_m(n), \tag{2}$$

where $h_m(n)$ is the room impulse response from the array center to the $m^{th}$ microphone, $s(n)$ is the test signal, and $u_m(n)$ is measurement and background noise. Given a persistently exciting signal $s(n)$ and an acceptable signal to noise ratio, one can estimate the room impulse responses (RIRs) from the observations $y_m(n)$. It is from these estimates that we infer the geometry of the room.

### C. Impulse response decomposition

We assume that the early reflections from an arbitrary RIR $h_m(n)$ may be approximately decomposed into a linear combination of the direct path and individual reflections, such that

$$h_m(n) = h_m^{(dp)}(n) + \sum_{i=1}^{R} \rho^{(i)} h_m^{(r_i,\theta_i,\phi_i)}(n) + v_m(n), \tag{3}$$

where $h_m^{(dp)}(n)$ is the direct path; $R$ is the total number of modeled reflections; the superscript $i$ is the reflection index; $h_m^{(r_i,\theta_i,\phi_i)}(n)$ is the WIR from a wall at position $(r_i, \theta_i, \phi_i)$, and from which the direct path from the loudspeaker to the microphone has been removed; $\rho^{(i)}$ is the reflection coefficient (which we assume to be frequency invariant); $v_m(n)$ includes noise, residual reflections and diffuse reverberation, which are not accounted in the summation.

Note that we assume that $\rho^{(i)}$ does not depend on $m$, and this claim deserves justification. While the reflection coefficient obviously depends on a wall and not on the array, it is conceivable (albeit unlikely) that the sound impinging on a pair of microphones could have reflected off different walls. However, for reasonably small arrays the sound will take approximately the same path from the source to each of the microphones, which implies that it should with high probability reflect off the same walls before reaching each microphone, such that the reflection coefficients will be the same for every microphone.

Now define

$$\begin{aligned}\mathbf{x}_m &= \left[\, x_m(0) \ \cdots \ x_m(N) \,\right]^T \\ \mathbf{x} &= \left[\, \mathbf{x}_1^T \ \cdots \ \mathbf{x}_M^T \,\right]^T \\ \mathbf{x}_{m,\tau} &= \left[\, x_m(\tau) \ \cdots \ x_m(N+\tau) \,\right]^T \\ \mathbf{x}_\tau &= \left[\, \mathbf{x}_{1,\tau}^T \ \cdots \ \mathbf{x}_{M,\tau}^T \,\right]^T\end{aligned}$$

for any signal $x_m(n)$ associated with the $m^{th}$ microphone.

We can then rewrite (3) in truncated vector form as

$$\mathbf{h} = \mathbf{h}^{(dp)}(n) + \sum_{i=1}^{R} \rho^{(i)} \mathbf{h}^{(r_i,\theta_i,\phi_i)} + \mathbf{v}, \tag{4}$$

where we have selected a vector length $N$ that is large enough to contain the 1st, 2nd and 3rd-order reflections, but that cuts off the higher order reflections and the reverberation tail. Therefore, given a measured $\mathbf{h}$, our problem is to estimate $\rho^{(i)}$ and $(r_i, \theta_i, \phi_i)$ for the dominant early reflections, which (after some post-processing) can reveal the position of the walls, floor and ceiling.

Our proposed method for room modeling first requires obtaining synthetically and/or experimentally for the array of interest a collection $\mathcal{H}_0 = \left\{ \underline{\mathbf{h}}^{(r_0,\theta,\phi)} \right\}_{\theta\in\Theta,\phi\in\Phi}$ of WIRs, each measured at fixed range $r = r_0$ over a grid $\Theta \subset [0, 360^\circ)$ of azimuth angles and $\Phi \subset [-90^\circ, 90^\circ]$ of elevation angles. We underline $\underline{\mathbf{h}}^{(r_0,\theta,\phi)}$ to highlight the fact that these WIRs are sampled over a discrete grid, at a single range, and model the reflective properties of a specific wall material.

In essence, $\mathcal{H}_0$ carries a time-domain description of the array manifold vector for multiple directions of arrival. If we assume a plane-wave approximation (valid from the small array assumption) and a sufficiently high sampling rate, given an arbitrary $\underline{\mathbf{h}}^{(r_*,\theta_*,\phi_*)}$ with $r_* > r_0$ we have that

$$\underline{\mathbf{h}}^{(r_*,\theta_*,\phi_*)} \approx \frac{r_0}{r_*} \underline{\mathbf{h}}_{\tau_*}^{(r_0,\theta_*,\phi_*)}, \tag{5}$$

for $\tau_* = \lfloor 2(r_* - r_0) \cdot f_s/c \rceil$, where $\lfloor \cdot \rceil$ denotes the nearest integer in samples, $f_s$ is the sampling rate and $c$ is the speed of sound. Thus, $\underline{\mathbf{h}}^{(r_0,\theta_*,\phi_*)}$ generates a family of reflections for a given direction. Since a room can be modeled as a linear system, if we assume that $\Theta \times \Phi$ is sufficiently fine, reflection coefficients are frequency-independent and we neglect the direct path from loudspeaker to microphone, any reflection can be expressed as a time-shifted and attenuated WIR. Thus, there are coefficients $\{c_i\}_{i=1}^R$ such that given an impulse response $\mathbf{h}_{room}$ which had the direct path removed and was truncated to only contain early reflections,

$$\mathbf{h}_{room} \approx \sum_{i=1}^{R} c_i \underline{\mathbf{h}}^{(r_i,\theta_i,\phi_i)}. \tag{6}$$

Thus, under the approximations above, $\mathcal{H}_0$ spans the space of truncated impulse responses which are measurable by a particular array.

Define $\mathcal{H}_\tau = \{\underline{\mathbf{h}}_\tau : \underline{\mathbf{h}} \in \mathcal{H}_0\}$ and $\mathcal{H}_* = \cup_{\tau=0}^{T} \mathcal{H}_\tau$, where $T$ is the maximum delay (in samples) we wish to model for a reflection. Our problem is then to fit elements $\mathcal{H}_*$ to the measured impulse response, adjusting for attenuation. A sparse solution is also required, given that $\mathcal{H}_*$ contains a very large number of candidate reflections, and we are interested in the dominant 1st, 2nd and 3rd-order reflections (from which the room geometry is inferred).

Consider an enumeration of $\mathcal{H}_0$ such that $\mathcal{H}_0 = \left\{\underline{\mathbf{h}}^{(1)}, ..., \underline{\mathbf{h}}^{(K)}\right\}$, with $K = |\mathcal{H}_0|$. Define

$$\underline{\mathbf{H}} = \left[ \underline{\mathbf{h}}_0^{(1)} \cdots \underline{\mathbf{h}}_T^{(1)} \cdots \underline{\mathbf{h}}_0^{(K)} \cdots \underline{\mathbf{h}}_T^{(K)} \right], \tag{7}$$

where each WIR appears for each integer sample delay $\tau$ such that $0 \leq \tau \leq T$. We then solve the following $\ell_1$-regularized least-squares minimization (also known as the LASSO problem [13], [14]):

$$\min_{\mathbf{a}} \|\mathbf{h}_{room} - \underline{\mathbf{H}}\mathbf{a}\|_2^2 \text{ subject to } \|\mathbf{a}\|_1 \leq \sigma, \tag{8}$$

where $\sigma$ controls the sparsity of the desired solution. In previous work [15] we used an alternative formulation known as basis pursuit with denoising (BPDN) in the Lagrangian form, given by

$$\min_{\mathbf{a}} \|\mathbf{h}_{room} - \underline{\mathbf{H}}\mathbf{a}\|_2^2 + \lambda \|\mathbf{a}\|_1, \tag{9}$$

where $\lambda$ is the regularization parameter. It can be shown that LASSO and BPDN are equivalent for appropriate choices of $\sigma$ and $\lambda$. LASSO has the advantage of providing a straightforward choice of $\sigma$, while more solvers are available for BPDN (e.g., [16]–[18]).

We only consider WIRs corresponding to the $L$ largest coefficients in $\mathbf{a}$, since most of the faint coefficients correspond to high order reflections or to clutter. To identify which of the large coefficients correspond to actual wall reflections and to extract the room geometry we implement a post-processing step, which is the topic the next subsection. $L$ should be sufficiently large to include enough 2nd and 3rd order reflections, which are used for wall validation. Excessive values of $L$ will unnecessarily increase computational cost, and increase the probability of false positives.

### *D. Wall validation*

By solving (8) with a suitable regularization parameter $\sigma$, one fits the measured impulse response with its dominant WIRs. However, real-world impulse responses feature numerous reflections created by obstacles such as people, furniture, light fixtures, counters and office equipment. In general, these reflectors appear because they are close to the array, but are not acoustically dominant for more distant sources. Furthermore, they usually cannot be well modeled as planar reflectors. While some reflectors can be modeled as rectangular reflectors of limited extension, their size cannot be estimated by our procedure. Thus, we choose to reject reflections generated by smaller objects, and consider only reflections generated by the walls, the floor and the ceiling.

To identify these specific reflections, we use the validation procedure outlined in the ValidReflectors function of Algorithm 1. We know from the image model that if a 1st-order reflection is caused by a wall, then its 2nd and 3rd-order reflections with the floor, ceiling and neighboring walls also exist. Thus, we only declare a reflector to be valid if at least one of its 2nd or 3rd order reflections are also detected. This discards reflections created by small objects.

Note that to validate a reflector candidate using a 2nd-order reflection, one must know a previously validated reflector. We bootstrap this method by assuming the ceiling corresponds to the strongest detectable vertical reflection. Indeed, in typical environments the ceiling is entirely visible to the microphones and to the loudspeaker, and is relatively close to the device. Thus, its 1st-order reflection is guaranteed to be very strong. Furthermore, only a small number of reflections arrive from $\phi \approx 90°$, and they will always involve the ceiling (the vast majority of reflections involve walls, and therefore have horizontal or shallow angles of arrival). Thus, the ceiling first-order reflection can be easily classified. By assuming that the ceiling is parallel to the floor, a (ceiling, floor) pair is considered valid if the 2nd-order image with respect to these two reflectors is also detected. We initialize the set $\mathcal{C}$ with all validated vertical reflections.

Due to unmodeled surfaces and frequency dependent reflection coefficients, the estimates for $\phi$ may not always be accurate. To avoid missing any first-order wall reflections (which would ideally have $\phi = 0°$), we initialize the set $\mathcal{W}$ of wall candidates with all reflections having $\phi < 35°$. We then test each wall candidate, and declare it to be valid if at least one 2nd or 3rd-order reflection (as predicted by the image model) was also detected.

In general, validation with wall-ceiling reflections delivers the lowest false positive probability. Indeed, wall-ceiling reflections are typically strong, since they only involve two reflectors and the ceiling is almost never occluded. Furthermore, due to the non-zero elevation, it is less likely for clutter to be incorrectly classified as a wall-ceiling reflection.

Due to the distances involved, wall-wall reflections are often fainter than wall-ceiling reflections, but still serve an important validation role. Nevertheless, it is possible for clutter to produce both a wall-candidate and a false wall-wall reflection which validates it, generating a false positive. On the other hand, each wall produces two wall-wall reflections (one for each neighbor), facilitating the detection of at least one of them.

Finally, wall-wall-ceiling reflections are the least reliable for validation. Even though their non-zero elevation is a desirable characteristic, their corresponding image sources can be distant enough to produce shallow angles of arrival, thus increasing the likelihood of occlusion. The interaction of three reflectors with unknown frequency responses can cause significant distortion to the pulse shapes, and the large propagation distances inevitably produce attenuation. Finally, the large number of 3rd-order reflections and the previous characteristics can make them difficult to distinguish from clutter.

**Algorithm 1** Reflector (wall, floor and ceiling) validation with image method constraints

// returns whether $\mathcal{R}_1$ and $\mathcal{R}_2$ have reflections in common
**function** haveCommonReflection($\mathcal{R}_1, \mathcal{R}_2$)
  **for** each $\mathbf{r}_1 \in \mathcal{R}_1$ and $\mathbf{r}_2 \in \mathcal{R}_2$ **do**
    **if** $\Delta(\mathbf{r}_1, \mathbf{r}_2) < \boldsymbol{\delta}$ **then**
      **return true**
  **return false**
**end function**

// returns whether $\mathbf{w}$, $\mathcal{C}$ and $\mathcal{W}$ generate a reflection in $\mathcal{R}$
**function** isValidWall($\mathbf{w}, \mathcal{C}, \mathcal{W}, \mathcal{R}$)
  $\mathcal{R}_{\text{wc}} = \{\Xi(\mathbf{w}, \bar{\mathbf{c}}) : \bar{\mathbf{c}} \in \mathcal{C}\}$
  $\mathcal{R}_{\text{ww}} = \{\Xi(\mathbf{w}, \bar{\mathbf{w}}) : \bar{\mathbf{w}} \in \mathcal{W}\}$
  $\mathcal{R}_{\text{wwc}} = \{\Xi(\mathbf{w}, \bar{\mathbf{w}}, \bar{\mathbf{c}}) : \bar{\mathbf{w}} \in \mathcal{W} \wedge \bar{\mathbf{c}} \in \mathcal{C}\}$
  $\mathcal{S} = \{\mathbf{r} \in \mathcal{R}_{\text{wc}} \cup \mathcal{R}_{\text{ww}} \cup \mathcal{R}_{\text{wwc}} : \mathbf{r} \text{ is feasible}\}$
  **return** haveCommonReflection($\mathcal{R}, \mathcal{S}$)
**end function**

// returns whether $\mathbf{c}$ and $\mathcal{C}$ generate a reflection in $\mathcal{R}$
**function** isValidFloorOrCeiling($\mathbf{c}, \mathcal{C}, \mathcal{R}$)
  $\mathcal{R}_{\text{cc}} = \{\Xi(\mathbf{c}, \bar{\mathbf{c}}) : \bar{\mathbf{c}} \in \mathcal{C}\}$
  $\mathcal{S} = \{\mathbf{r} \in \mathcal{R}_{\text{cc}} : \mathbf{r} \text{ is feasible}\}$
  **return** haveCommonReflection($\mathcal{R}, \mathcal{S}$)
**end function**

// returns whether $\mathbf{w}$ is not generated by $\mathcal{C}$ and $\mathcal{W}$
**function** isNotReflector($\mathbf{w}, \mathcal{C}, \mathcal{W}$)
  $\mathcal{R}_{\text{wc}} = \{\Xi(\mathbf{w}_1, \mathbf{c}) : \mathbf{w}_1 \in \mathcal{W} \wedge \mathbf{c} \in \mathcal{C}\}$
  $\mathcal{R}_{\text{ww}} = \{\Xi(\mathbf{w}_1, \mathbf{w}_2) : \mathbf{w}_1 \in \mathcal{W} \wedge \mathbf{w}_2 \in \mathcal{W}\}$
  $\mathcal{R}_{\text{wwc}} = \{\Xi(\mathbf{w}_1, \mathbf{w}_2, \mathbf{c}) : \mathbf{w}_1 \in \mathcal{W} \wedge \mathbf{w}_2 \in \mathcal{W} \wedge \mathbf{c} \in \mathcal{C}\}$
  $\mathcal{S} = \{\mathbf{r} \in \mathcal{R}_{\text{wc}} \cup \mathcal{R}_{\text{ww}} \cup \mathcal{R}_{\text{wwc}} : \mathbf{r} \text{ is feasible}\}$
  **return** $\neg$haveCommonReflection($\{\mathbf{w}\}, \mathcal{S}$)
**end function**

// $\mathcal{R}$ : reflections corresponding to the $L$ largest
// coefficients from $\mathbf{a}$, sorted from strongest to weakest
// returns the valid reflectors (walls, floor and ceiling)
**function** ValidReflectors($\mathcal{R}$)
  $\mathcal{C} = \{(r, \theta, \phi) \in \mathcal{R} : \phi = 90°\}$
  $\mathbf{c}_1 =$ strongest reflection in $\mathcal{C}$ // presumed ceiling
  $\mathcal{C} = \{\mathbf{c} \in \mathcal{C} : \text{isValidFloorOrCeiling}(\mathbf{c}, \{\mathbf{c}_1\}, \mathcal{R})\}$
  $\mathbf{c}_2 =$ strongest reflection in $\mathcal{C}$ // presumed floor
  $\mathcal{C} = \{\mathbf{c}_1, \mathbf{c}_2\}$
  $\mathcal{W} = \{(r, \theta, \phi) \in \mathcal{R} : \phi < 35°\}$ // wall candidates
  $\mathcal{W} = \{\mathbf{w} \in \mathcal{W} : \text{isValidWall}(\mathbf{w}, \mathcal{C}, \mathcal{W}, \mathcal{R})\}$
  $\mathcal{W} = \{\mathbf{w} \in \mathcal{W} : \text{isNotReflector}(\mathbf{w}, \mathcal{C}, \mathcal{W})\}$
  $\mathcal{W} = \text{merge}(\mathcal{W}, \boldsymbol{\delta})$ // see Algorithm 2
  **return** $\mathcal{C} \cup \mathcal{W}$
**end function**

Thus, a practical implementation should annotate each wall with the order of the reflections used for its validation, since these indicate the confidence in the produced model. In our MATLAB implementation, if there is more than one wall candidate for a given direction of arrival, the orders of the validating reflections are used as tie-breakers.

**Algorithm 2** Reflection merging function

// returns the mean of vectors in spherical coordinates
**function** sphericalMean($\mathcal{R}$)
  $\mathcal{R} = \text{sph2cart}(\mathcal{R})$ // spherical to cartesian transform
  $\mathbf{r} = \text{mean}(\mathcal{R})$ // mean vector
  $\mathbf{r} = \text{cart2sph}(\mathbf{r})$ // cartesian to spherical transform
  **return** $\mathbf{r}$
**end function**

// merges close reflection normals into their mean
**function** merge($\mathcal{R}, \boldsymbol{\delta}$)
  $\mathcal{S} = \emptyset$
  **for** each $\mathbf{r} \in \mathcal{R}$ **do**
    $\mathcal{M} = \{\mathbf{m} \in \mathcal{R} : \Delta(\mathbf{m}, \mathbf{r}) < \boldsymbol{\delta}\}$
    $\mathcal{R} = \mathcal{R} \backslash \mathcal{M}$
    $\mathcal{S} = \mathcal{S} \cup \{\text{sphericalMean}(\mathcal{M})\}$
  **return** $\mathcal{S}$
**end function**

Note that since $\mathcal{W}$ was initialized with all shallow reflections, it also contains 2nd and 3rd-order reflections. These can be incorrectly validated as 1st-order reflections (we present examples in Section V-B). The isNotReflector test (see Algorithm 1) is designed to discard these false positives.

Finally, it is possible to refine the 1st-order reflections in $\mathcal{C} \cup \mathcal{W}$ to obtain more accurate coordinates. Let $\mathcal{C} \cup \mathcal{W} = \{\bar{\mathbf{r}}_1, \cdots, \bar{\mathbf{r}}_N\}$. The refinement consists of solving

$$\min_{\mathbf{a}, \mathbf{r}_1, \cdots, \mathbf{r}_N} \|\mathbf{h}_{room} - \underline{\mathbf{H}}^{\mathbf{r}} \mathbf{a}\|_2^2 , \tag{10}$$

with

$$\underline{\mathbf{H}}^{\mathbf{r}} = \left[ \underline{\mathbf{h}}^{\mathbf{r}_1} \ \cdots \ \underline{\mathbf{h}}^{\mathbf{r}_N} \right]$$
$$\mathbf{r}_i = \bar{\mathbf{r}}_i + (r_i, \theta_i, 0°)$$

subject to $|r_i| < \delta_r$ and $|\theta_i| < \delta_\theta$. In practice, to simplify the solution of (10), we decouple $\mathbf{r}_1, \cdots, \mathbf{r}_N$ and optimize one $\mathbf{r}_i$ at a time.

## IV. Practical considerations

### A. WIR acquisition and array modeling

This proposed method relies on the knowledge of $\underline{\mathbf{h}}^{(r_0, \theta_i, \phi_i)}$, over a grid of azimuth and elevation angles. The distance $r_0$ must be sufficiently large with respect to the array size, so that the far-field approximation (5) is valid. For arrays of omnidirectional microphones with acoustically transparent enclosures and well defined geometries, WIRs may be determined analytically to good tolerance, given the loudspeaker's transfer function and the gains for all microphones and associated conditioning circuits.

In practice, one may encounter designs involving characteristics which are difficult to model without laboratory measurements. In our experiments, we used the RoundTable device, which is a 6-element circular array of cardioid microphones. The microphones are housed in a plastic enclosure which further shapes their spatial patterns. The RoundTable's integrated loudspeaker is intended for teleconferencing, and features a colored frequency response. To produce an accurate

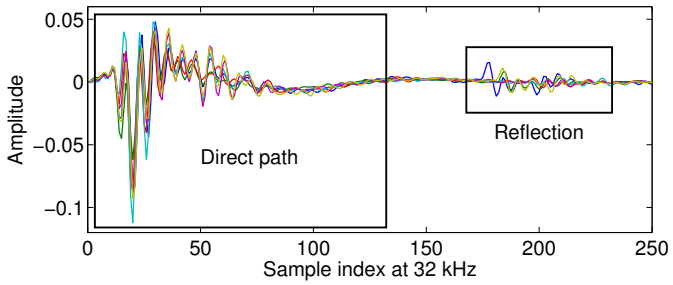



Figure 4: RIR for WIR at $\theta = \phi = 0°$.

model, we developed a methodology for reliably acquiring WIRs and generating an array model, which we describe in this section.

WIRs were collected experimentally in an anechoic chamber, using a large 1" thick acrylic barrier as a wall simulator. The array was attached to a custom built mount, which was fit to a tripod capable of swiveling in azimuth and elevation. The integrated loudspeaker was used to generate 1-second sine sweeps, from which impulse responses were estimated for every $15°$ in azimuth and $10°$ in elevation, for a total of 240 DOAs.

Fig. 4 shows impulse responses for $\theta = \phi = 0°$, for all 6 microphones. The integrated loudspeaker's response is far from ideal, and also presents acoustic coupling with the enclosure, leading to the highlighted direct path. The reflection from the acrylic barrier can be easily extracted, since it appears after the tail of the direct path, and the range to the barrier is known.

Since the WIRs featured in $\underline{\mathbf{H}}$ span more DOAs than the 240 experimentally sampled values, we use bilinear interpolation in azimuth and elevation to synthesize WIRs for arbitrary DOAs. Since each channel has a DOA-dependent delay, we interpolate between time-aligned templates, and delay them with subsample accuracy to simulate the propagation delay and phase shift applicable to each microphone. These templates are intended to be device-independent, with the exception of a microphone-specific gain which must be estimated during manufacturing (as is already performed for every RoundTable device). A consumer device designed for this application should use high quality MEMS microphones, to minimize the probability of manufacturing a mismatched unit. By sourcing quality components, the greatest source of impulse response mismatch should be due to the unkown reflector materials and sizes, and not due to microphone variations.

To generate a set of WIR templates suitable for interpolation, we:

1) Resample the impulse responses from 16 kHz to 32 kHz. This allows us to double the range accuracy when solving (8), since $\underline{\mathbf{H}}$ only features WIRs with integer sample delays. It also allows us to apply fractional delay filters without distorting the WIRs[1].
2) Use a Lagrange fractional delay filter [19] to find the subsample delay needed to maximize the peak of the reflection for the closest microphone. This aligns the strongest reflection with subsample accuracy at a known temporal coordinate, which is important since the 3D coordinates of the image loudspeaker with respect to the array are not exactly known (due to non-systematic alignment errors related to the tripod mount).
3) Use the array geometry to time align all other microphones, undoing the relative propagation delays with subsample accuracy.
4) Extract the 20 sample[2] long section of the RIR corresponding to the reflection, creating a reflection template for every measured DOA.
5) For all directions of arrival in the measurement grid, replace the templates by averages over all microphones that have the same (azimuth, elevation) pair, under the assumption that microphone responses are symmetric in azimuth (see Fig. 5).
6) For all directions of arrival in the measurement grid, estimate the delay between the closest microphone and all other microphones, using the cross-correlation peak with subsample resolution. Store this delay, and align all microphones.
7) Repeat the averaging from step (5), now with the aligned templates (see Fig. 5).

[1]Lagrange fractional delay filters [19] have maximally flat gain around DC, but are band-limited. By upsampling the RIRs, we guarantee that Lagrange filters will not cause signal distortion, since sufficiently long filters have nearly constant group delay for frequencies under 1/2 of the Nyquist frequency.

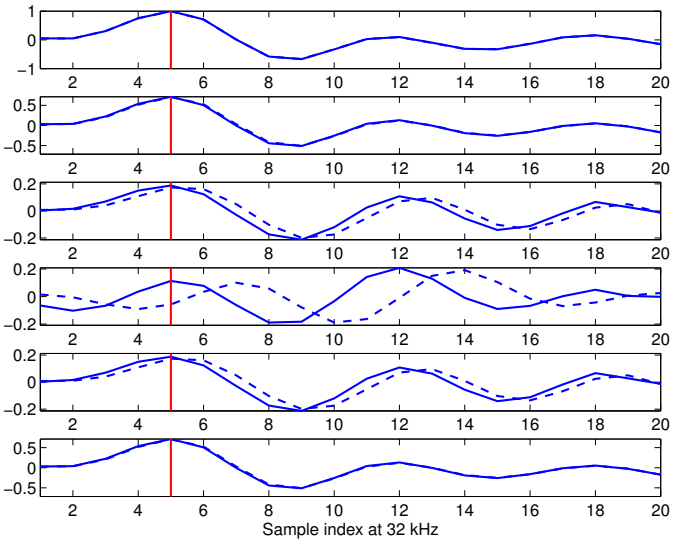



Figure 5: Templates for the WIR at $\theta = \phi = 0°$ for microphones 1-6, with the maximum for the closest microphone at sample 5. Dashed line: averaged templates, before phase alignment (step 5), continuous line: averaged, with phase alignment (step 7).

Cardioid microphones have azimuth-dependent phase shifts, with a maximum for $\theta = 180°$. Step (6) estimates this phase shift as a delay, which is re-applied when creating WIRs for arbitrary DOAs. By averaging the measured WIRs before step (6), we mitigate measurement errors incurred from uncertainties in azimuth and elevation.

To generate a WIR for an arbitrary DOA, we use bilinear

[2]When sampling at 16 kHz, experiments showed that for a wide variety of surfaces, 20 samples are sufficient to capture the peak and speaker-related oscillations of a reflection, without the risk of overfitting the measurement.

interpolation between the four templates with neighboring azimuth and elevation values. This creates an approximate template for the desired DOA. For each microphone, we then use bilinear interpolation for the four delays with neighboring azimuth and elevation values. These delays are applied to each channel, along with the geometric delay computed using the source and array locations, producing the synthetic WIR.

### B. Transform implementation

A practical consideration involves the computational tractability of solving (8). While the desirable range resolution is application dependent, a rule of thumb is to require range resolution corresponding to the propagation delay of 1 sample at the desired sampling frequency. This ensures that synthetically generated early reflections will be typically aligned within 1 sample of the ground truth.

For example, consider a sampling rate $f_s = 16\,\mathrm{kHz}$ and let $c = 345\,\mathrm{m/s}$ be the speed of sound. If one wishes to identify walls located between 1.0 and 7.0 meters from the array, one must plan for a round-trip time varying between $2 \cdot 1.0 \cdot \frac{f_s}{c} \approx 93$ and $2 \cdot 7.0 \cdot \frac{f_s}{c} \approx 649$ sample delays, which implies $T = 649 - 93 + 1 = 557$. The grid of single wall reflections should be sufficiently fine, otherwise reflections may be incorrectly identified. By sampling in azimuth with $2^\circ$ resolution and in elevation with $10^\circ$ resolution, we have $K = 1621$ WIRs (180 WIRs from each elevation angle from $0^\circ$ to $80^\circ$, plus 1 WIR for an elevation of $90^\circ$). Therefore, $\underline{\mathbf{H}}$ has $T \cdot K = 902897$ columns. A $7\,\mathrm{m} \times 7\,\mathrm{m} \times 2\,\mathrm{m}$ room will have a 3rd-order reflection generated from a virtual source at a distance of $2\sqrt{7^2 + 7^2 + 2^2} \approx 20.0\,\mathrm{m}$. To model this reflection, we must consider an impulse response with at least $20.0\frac{f_s}{c} \approx 936$ samples. For an array with 6 microphones, an explicit representation of $\underline{\mathbf{H}}$ becomes a $5616 \times 902897$ matrix, which is too large to be explicitly represented when solving (8).

To solve (8) using sparse-recovery algorithms such as [14], [16] one must implement the $\underline{\mathbf{H}}\mathbf{x}$ and $\underline{\mathbf{H}}^T\mathbf{y}$ matrix-vector products for arbitrary $\mathbf{x}$ and $\mathbf{y}$. Fortunately, it is possible to exploit $\underline{\mathbf{H}}$'s block matrix nature to avoid representing $\underline{\mathbf{H}}$ explicitly, and also to accelerate the products. Indeed, $\underline{\mathbf{H}}$ can be written as

$$\underline{\mathbf{H}} = \left[\, \underline{\mathbf{H}}^{(1)} \; \underline{\mathbf{H}}^{(2)} \; \cdots \; \underline{\mathbf{H}}^{(K)} \,\right], \tag{11}$$

where

$$\underline{\mathbf{H}}^{(i)} = \left[\, \underline{\mathbf{h}}^{(i)}_{\tau=0} \; \underline{\mathbf{h}}^{(i)}_{\tau=1} \; \cdots \; \underline{\mathbf{h}}^{(i)}_{\tau=T} \,\right]. \tag{12}$$

It is easy to see that for all $i$, $\underline{\mathbf{H}}^{(i)}$ is Toeplitz. Therefore, $\underline{\mathbf{H}}^{(i)}\mathbf{x} = \underline{\mathbf{h}}^{(i)}_{\tau=0} * \mathbf{x}$ (where $*$ represents linear convolution, truncated to the length of $\underline{\mathbf{h}}^{(i)}_{\tau=0}$). Using a small amount of zero-padding (since each $\underline{\mathbf{h}}^{(i)}_{\tau=0}$ has a very small support), this can be accelerated with an FFT. It is easy to show that $\left[\underline{\mathbf{H}}^{(i)}\right]^T \mathbf{y} = \underline{\mathbf{h}}^{(i)}_{\tau=0} \star \mathbf{y}$ (where $\star$ denotes cross-correlation), which can also be evaluated with FFTs. Using this method, both matrix-vector products can be performed using $K$ fast convolutions or fast correlations.

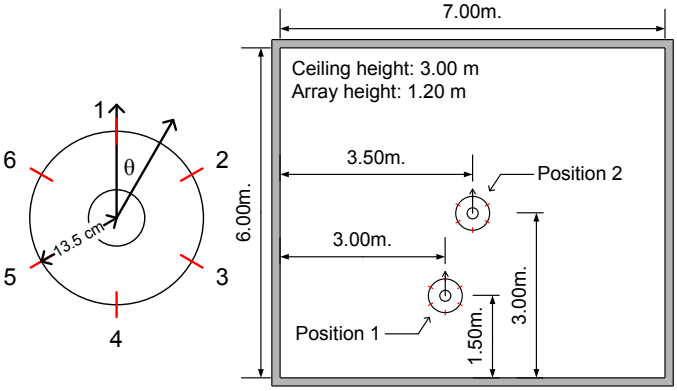



Figure 6: Array geometry and synthetic room dimensions.

Finally, to ensure that certain DOAs will not be favored over others, we normalize all WIRs such that

$$\left\|\underline{\mathbf{h}}^{(1)}_{\tau=0}\right\|_2 = \left\|\underline{\mathbf{h}}^{(2)}_{\tau=0}\right\|_2 = \cdots = \left\|\underline{\mathbf{h}}^{(K)}_{\tau=0}\right\|_2 = 1.$$

## V. Results

### A. Image model simulations

Next we present image model [3] simulations designed to illustrate the proposed method. Fig. 6 shows the simulated array geometry, designed to coincide with the RoundTable device used in the experiments. The same figure also shows the simulated room, which measures $6 \times 7 \times 3\,\mathrm{m}$ and has a reverberation time of 300 ms. The walls, floor and ceiling have a frequency independent reflection coefficient of 0.77.

The synthetic impulse response was estimated using a 30 Hz to 8 kHz linear sine sweep, using frequency domain division [20]. This estimation was performed under an SNR of 20 dB, generated by adding Gaussian white noise. Note that while this SNR is compatible with real-world measurements using consumer equipment designed for teleconferencing, this scenario is very favorable, since we do not simulate clutter, assume the walls to be perfectly flat and the array model to be exact.

The array was simulated considering cardioid microphones with a spatial gain response given by $[\cos(\theta) + 1.1]/2.1$, where $\theta$ is the angle between the microphone's axis and the DOA. The synthetic loudspeaker and microphones were simulated with a flat frequency response, and zero phase-shifts. $\underline{\mathbf{H}}$ was generated with WIRs at $(1.0, \theta, \phi)$, with $\theta \in \{0^\circ, 2^\circ, \cdots, 358^\circ\}$ and $\phi \in \{0^\circ, 10^\circ, \cdots, 90^\circ\}$, and delayed for every integer sample delay (as prescribed in Section IV-B) for $1.0 \le r \le 7.0\,\mathrm{m}$. The LASSO minimization (8) was solved with $\sigma = 5$ using the solver SPGL1 [14]. The array resolution was set to $\boldsymbol{\delta} = (.05, 10^\circ, 25^\circ)$.

Fig. 7 shows annotated impulse responses for position 1, plotted without noise to facilitate the visualization. We highlight the 1st-order reflections corresponding to each of the 4 walls, for each of the 6 microphones. We also highlight higher order (typically 2nd and 3rd-order) reflections. The first version of this method [15] only considered 1st-order reflections and a WIR model featuring only $\phi = 0^\circ$ and $\phi = 90^\circ$. This simplified method can be used by assuming the walls make $90^\circ$ angles with each other. Nevertheless, it is

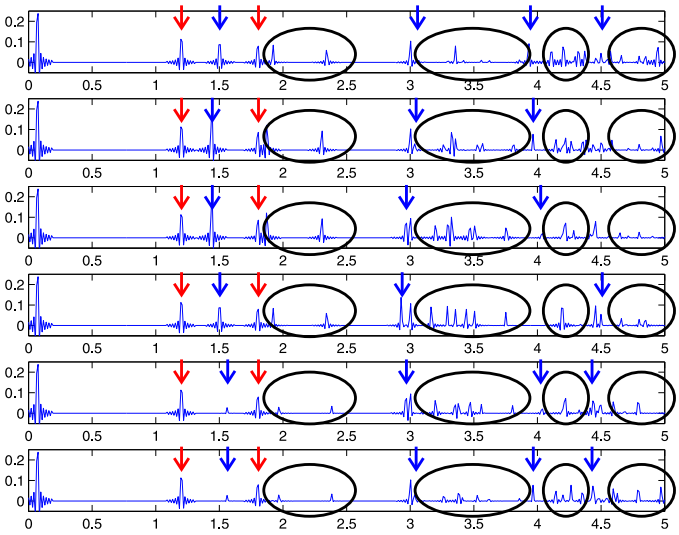


Figure 7: Synthetic RIRs for position 1. Blue arrows indicate 1st-order reflections with the walls, red arrows indicate 1st-order reflections with the floor and ceiling, and ellipses indicate higher order reflections.

Table I: Estimated walls for the synthetic room at position 1.

| Ground Truth | | | Estimates | | |
|---|---|---|---|---|---|
| $r$ (m) | $\theta$ (°) | $\phi$ (°) | $r$ (m) | $\theta$ (°) | $\phi$ (°) |
| 1.200 | 0 | -90 | 1.200 | 0 | 90 |
| 1.800 | 0 | 90 | 1.800 | 0 | 90 |
| 4.500 | 0 | 0 | 4.502 | 0 | 0 |
| 4.000 | 90 | 0 | 4.000 | 90 | 0 |
| 1.500 | 180 | 0 | 1.500 | 180 | 0 |
| 3.000 | 270 | 0 | 3.005 | 270 | 0 |

not as reliable as this current proposal. From Fig. 7 it should be clear that 2nd and 3rd-order reflections are numerous, and should be modeled to increase the method's sensitivity without producing a prohibitive number of false positives.

Tables I and II show results for positions 1 and 2. In both cases, all walls are identified within a few mm of the ground truth. Note that wall identification at position 2 is more challenging, since each pair of opposing walls is at exactly the same range. Nevertheless, all walls were estimated correctly. Since the circular array cannot discriminate between floor and ceiling, the floor is identified as a horizontal reflector (which we represent with $\phi = 90°$).

To investigate performance over more scenarios, we simulated all source locations parameterized by $(x, y) = (t, t)$, for $t \in \{1.0, 1.1, \cdots, 5.0\}$. All surfaces were estimated at the correct range (with an error under 1 cm) for all cases, except for $t \leq 1.4$, when the wall at $0°$ could not be detected. Note that when $t \leq 1.4$, this wall is farther than 5.6 m from the array. The fast implementation of $\underline{\mathbf{H}}$ only shifts WIRs, without modeling the energy decay due to distance. Thus, distant walls are assigned small coefficients in the $\ell_1$-regularized least-squares procedure, and may not be detected.

Table II: Estimated walls for the synthetic room at position 2.

| Ground Truth | | | Estimates | | |
|---|---|---|---|---|---|
| $r$ (m) | $\theta$ (°) | $\phi$ (°) | $r$ (m) | $\theta$ (°) | $\phi$ (°) |
| 1.200 | 0.0 | -90 | 1.200 | 0.0 | 90 |
| 1.800 | 0.0 | 90 | 1.800 | 0.0 | 90 |
| 3.000 | 0.0 | 0 | 3.001 | 0.0 | 0 |
| 3.500 | 90.0 | 0 | 3.501 | 90.0 | 0 |
| 3.000 | 180.0 | 0 | 3.001 | 180.0 | 0 |
| 3.500 | 270.0 | 0 | 3.501 | 270.0 | 0 |

### B. *Real conference rooms*

Experimental data were captured using a RoundTable device (RTD), which features a 6-microphone array with the geometry illustrated in Fig. 1. Its microphones are cardioid, and the device features a loudspeaker in its center. Since the loudspeaker is intended for teleconferencing applications, it features distortions and nonlinearities that would not be present when using reference loudspeakers.

In our experiments, the RTD was always set on a large conference room table. Note that the presence of the table prevents the detection of the floor. Since the RTD microphones are mounted flush to its base, reflections from the table cannot be detected either. Thus, the following room models only feature walls and the ceiling.

Using an anechoic chamber and an RTD, we obtained WIRs with a resolution of $15°$ in azimuth and $10°$ in elevation, for a total of 240 DOAs. All impulse responses were estimated using a 1-second linear sweep from 30 Hz to 8 kHz, reproduced through the RTD's integrated speaker. Templates were extracted as prescribed in Section IV-A. The remaining parameters are the same as used for the simulations, except for $\sigma = 0.5$ (which differs from the value used for simulations because the amplitudes of the experimental signals are relatively small). As is the case with the simulations, results are not sensitive to the choice of $\sigma$.

By analyzing the templates, it becomes apparent that the RoundTable is not the ideal device to capture reflections coming from walls. Indeed, its microphone enclosures were designed to deliver the highest gain to signals arriving from around $\phi = 30°$, to attenuate the contribution of reflections and reverberation. Additionally, the RoundTable loudspeaker is mounted facing upwards, such that its directivity is low to the sides. Thus, some secondary reflections from the ceiling and walls are often detected with better clarity than the primary reflections from the side walls. In particular, the reflection from the ceiling is very strong. Its detection is also favored, since it corresponds to the only WIR where the reflection is not delayed between microphones. Thus, one can reliably assume the ceiling to be the strongest reflection with $\phi = 90°$.

Impulse responses were collected from 9 conference rooms, which were fully equipped and decorated. Their floorplans and estimation results are shown in Fig. 8. The wall coordinates and reflection orders used for validation are summarized in Table III. The ranges to the walls were measured with a laser range finder. The distances in parentheses correspond to estimates produced by the proposed method. All rooms have a rectangular floorplan, although slight deviations from orthogonality (on the order of $1°$ or less) are normal. Thus, we do not annotate the azimuth angle for the ground truth. In all rooms, the array was visually aligned with a wall at $0°$. While visual alignment is remarkably accurate, small deviations can be expected.

The Refs column of Table III indicates which reflections were used to validate each wall. This field has the format ABC, where A, B and C indicate the number of wall-ceiling, wall-wall and wall-wall-ceiling reflections used for validation. Thus, $A \in \{0,1\}$, $B \in \{0,1,2\}$ and $C \in \{0,1,2\}$. Entries with dashes (–) are present for walls which were not detected, or for the ceiling, which requires no validation.

Most of the conference rooms in this study feature walls of differing materials. Each conference room typically has one smooth wall reserved as a projection surface, a second wall with a whiteboard starting at table height, and a fabric covered bulletin board covering at least one of the remaining walls. For example, Fig. 9 shows the panorama corresponding to the room from Fig. 8f.

We note that in most cases, the wall panels begin approximately at table height, but may not extend all the way to the ceiling. Thus, the detected 1st-order reflection produces the distance to the panel, and not to the wall behind it. On the other hand, when the panel does not extend all the way to the ceiling, wall-ceiling reflections consider the distance the underlying wall. This discrepancy may be accounted for with a suitable choice of $\boldsymbol{\delta}$. In these experiments, we used $\boldsymbol{\delta} = (.05\,\mathrm{m}, 10°, 25°)$ with good results.

In all floorplans from Fig. 8, 2nd and 3rd-order reflections are annotated with checkmarks (✓) if they properly validate real walls and Xs (✗) if they mistake a high-order reflection for a real wall, creating a false-positive. Corner-reflections are annotated with WW if they correspond to wall-wall 2nd-order reflections or WWC if they correspond to wall-wall-ceiling 3rd-order reflections.

Note that whenever a wall is close, its 2nd-order reflection with the ceiling tends to be detectable, and is responsible for validating the wall. All reflections from distant walls are faint, such that they risk being crowded out by closer walls. Nevertheless, each wall can be validated by up to four 2nd and 3rd-order reflections with its immediate neighbors, and usually at least one of them is detected.

The false-positive obtained in Fig. 8b represents a case where a WC reflection was fit with a shallow angle of arrival, mistaken for 1st-order reflection, and incorrectly validated by a WWC reflection (which in turn was identified as a WC reflection). Indeed, note that $\sqrt{2.71^2 + 2.00^2} \approx 3.37$, such that the modeled wall has the range of the WC reflection. This is consistent with the image model, because the 1st-order reflection associated with the wall at $\theta = 180°$ was not detected and a rectangular room assumption is not enforced. Thus, the isNotReflector test from Algorithm 1 was unable to reject this WC reflection.

One can reduce the occurrence of this type of false positive by not using WWC reflections for validation. However, we would then miss reflectors such as the wall at $\theta = 90°$ from Fig. 8f, which was only validated using a WWC reflection.

A related false positive is the wall at $\theta = 30°$ shown in Fig. 8c. It is in fact a WW reflection which was validated by a WWC reflection. Since the real wall at $\theta = 0°$ could not be detected, this 2nd-order reflection was incorrectly classified as a wall and could also not be rejected by the isNotReflector test from Algorithm 1.

Table III: Estimated walls for real rooms.

| Room | Ground Truth | | | Estimates | | | |
|---|---|---|---|---|---|---|---|
| | $r$ (m) | $\theta$ (°) | $\phi$ (°) | $r$ (m) | $\theta$ (°) | $\phi$ (°) | Refs |
| A | 1.81 | 0 | 90 | 1.81 | 0 | 90 | – |
| | 1.40 | 0 | 0 | 1.41 | 3 | 0 | 111 |
| | 2.11 | 90 | 0 | 2.11 | 94 | 0 | 111 |
| | 2.36 | 180 | 0 | 2.37 | 181 | 0 | 100 |
| | 3.91 | 270 | 0 | 3.92 | 273 | 0 | 100 |
| B | 2.00 | 0 | 90 | 2.00 | 0 | 90 | – |
| | 2.93 | 0 | 0 | 2.93 | 358 | 0 | 101 |
| | 2.46 | 90 | 0 | 2.46 | 92 | 0 | 100 |
| | 2.71 | 180 | 0 | 3.36 | 173 | 0 | 001 |
| | 1.82 | 270 | 0 | 1.83 | 270 | 0 | 101 |
| C | 2.00 | 0 | 90 | 2.00 | 0 | 90 | – |
| | 3.63 | 0 | 0 | – | – | – | – |
| | – | – | – | 3.94 | 30 | 0 | 100 |
| | 1.55 | 90 | 0 | 1.55 | 90 | 0 | 101 |
| | 2.02 | 180 | 0 | 2.02 | 182 | 0 | 102 |
| | 2.37 | 270 | 0 | 2.38 | 271 | 0 | 101 |
| D | 1.82 | 0 | 90 | 1.82 | 0 | 90 | – |
| | 2.75 | 0 | 0 | 2.75 | 4 | 0 | 011 |
| | 3.20 | 90 | 0 | – | – | – | – |
| | 4.46 | 180 | 0 | 4.47 | 184 | 0 | 100 |
| | 3.89 | 270 | 0 | 3.88 | 271 | 0 | 101 |
| E | 1.80 | 0 | 90 | 1.80 | 0 | 90 | – |
| | 4.55 | 0 | 0 | 4.55 | 0 | 0 | 111 |
| | 3.26 | 90 | 0 | – | – | – | – |
| | 2.65 | 180 | 0 | 2.65 | 178 | 0 | 111 |
| | 3.83 | 270 | 0 | 4.26 | 271 | 0 | 022 |
| F | 1.81 | 0 | 90 | 1.82 | 0 | 90 | – |
| | 2.23 | 0 | 0 | 2.24 | 0 | 0 | 101 |
| | 2.01 | 90 | 0 | 2.01 | 89 | 0 | 001 |
| | 1.49 | 180 | 0 | 1.49 | 180 | 0 | 110 |
| | 3.35 | 270 | 0 | 3.36 | 273 | 0 | 010 |
| G | 2.00 | 0 | 90 | 2.00 | 0 | 90 | – |
| | 2.36 | 0 | 0 | 2.34 | 1 | 0 | 111 |
| | 2.47 | 90 | 0 | 2.47 | 90 | 0 | 101 |
| | 3.29 | 180 | 0 | 3.30 | 188 | 0 | 101 |
| | 1.45 | 270 | 0 | 1.46 | 274 | 0 | 111 |
| H | 2.03 | 0 | 90 | 2.02 | 0 | 90 | – |
| | 4.34 | 0 | 0 | 4.33 | 359 | 0 | 010 |
| | 2.28 | 90 | 0 | 2.29 | 91 | 0 | 110 |
| | 2.62 | 180 | 0 | 2.62 | 178 | 0 | 121 |
| | 2.87 | 270 | 0 | 2.86 | 270 | 0 | 121 |
| I | 2.04 | 0 | 90 | 2.04 | 0 | 90 | – |
| | 3.08 | 0 | 0 | 3.09 | 3 | 0 | 101 |
| | 5.27 | 90 | 0 | – | – | – | – |
| | 2.26 | 180 | 0 | 2.26 | 181 | 0 | 101 |
| | 2.39 | 270 | 0 | 2.38 | 274 | 0 | 102 |

The column in Fig. 8e provides a difficult case, which was correctly identified using the WC reflection between the column and the ceiling. This is an unusual case, since columns are not common obstacles in conference rooms. Furthermore, while a column can have a large cross-section to a nearby source, it is not a major reflector with respect to most source locations. Still, it fits the heuristics of our model, and was validated as intended.

Most walls are identified correctly, with a typical range resolution of 1 cm. While some distant walls are not identified, they are in the minority, and do not significantly impact the model of early reflections.

## VI. Conclusion

We have presented a method for reliably obtaining room models using a small microphone array with an integrated loudspeaker. This information can be used in many acoustic

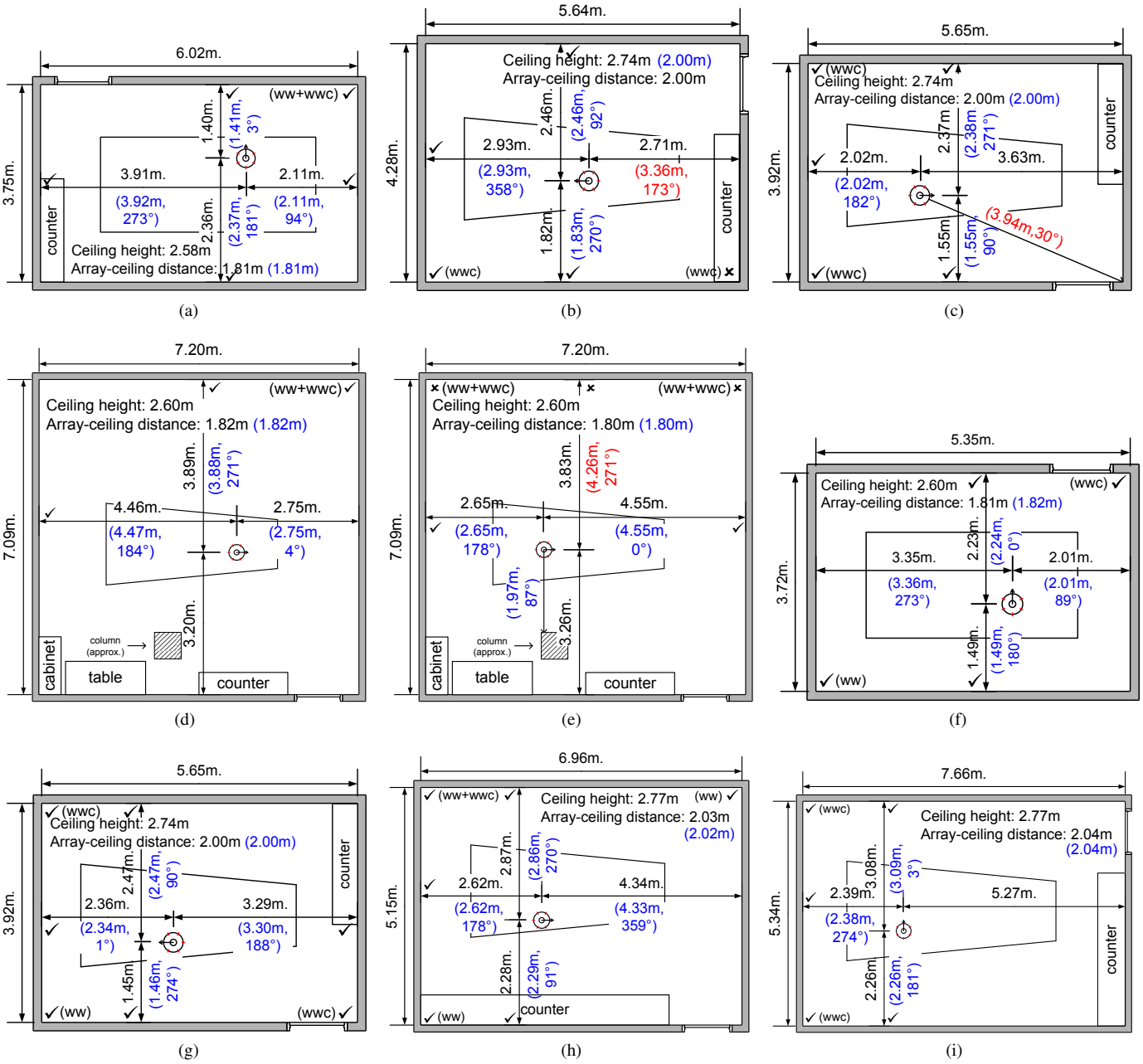



Figure 8: Experimental results obtained in real conference rooms.

Figure 9: Panorama for the room illustrated in Fig. 8f.

signal processing applications, including sound source localization, sound field reproduction and beamforming. We expect it will also open new research opportunities in acoustics, which will further explore the role of early reflections.

The proposed method uses a time-domain device model to obtain the strongest reflections which best fit a set of room impulse responses. Even though these reflections are numerous and may be distorted by characteristics which are generally unknown (such as clutter and the frequency-dependent reflection coefficients of materials), they contain the locations of the walls. By exploiting the sparsity of early reflections with respect to the space of all possible reflections, we obtain a stable and robust means of producing wall candidates. By enforcing the structural constraints embedded in secondary reflections, we validate wall candidates, thus producing a room model.

Experimental results show consistently good results, with a typical accuracy of 1 cm. The tests were performed using a 6-element off-the-shelf microphone array, in real corporate environments.

## REFERENCES


[1] F. Ribeiro, C. Zhang, D.A. Florêncio, and D.E. Ba, "Using reverberation to improve range and elevation discrimination for small array sound source localization," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 7, pp. 1781–1792, 2010.

[2] M.S. Song, C. Zhang, D. Florêncio, and H.G. Kang, "Enhancing loudspeaker-based 3D audio with room modeling," in *Proc. MMSP*, 2010.

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

[4] F. Remondino and S. El-Hakim, "Image-based 3D modeling: a review," *The Photogrammetric Record*, vol. 21, no. 115, pp. 269–291, 2006.

[5] M. Moebus and A.M. Zoubir, "Three-Dimensional Ultrasound Imaging in Air using a 2D Array on a Fixed Platform," in *Proc. of ICASSP*, 2007.

[6] A. O'Donovan, R. Duraiswami, and D. Zotkin, "Imaging concert hall acoustics using visual and audio cameras," in *Proc. of ICASSP*, 2008, pp. 5284–5287.

[7] D. Aprea, F. Antonacci, A. Sarti, and S. Tubaro, "Acoustic reconstruction of the geometry of an environment through acquisition of a controlled emission," in *Proc. of EUSIPCO*, 2009.

[8] F. Antonacci, A. Sarti, and S. Tubaro, "Geometric reconstruction of the environment from its response to multiple acoustic emissions," in *Proc. of ICASSP*, 2010.

[9] 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 *Proc. of ICASSP*, 2011.

[10] I. Dokmanic, Y.M. Lu, and M. Vetterli, "Can one hear the shape of a room: the 2-D polygonal case," in *Proc. of ICASSP*, 2011.

[11] J. Filos, E.A.P. Habets, and P.A. Naylor, "A Two-Step Approach to Blindly Infer Room Geometries," in *Proc. of the International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2010.

[12] J. Borish, "Extension of the image model to arbitrary polyhedra," *The Journal of the Acoustical Society of America*, vol. 75, pp. 1827, 1984.

[13] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[14] E. van den Berg and M.P. Friedlander, "Probing the Pareto frontier for basis pursuit solutions," *SIAM Journal on Scientific Computing*, vol. 31, no. 2, pp. 890–912, 2008.

[15] D. Ba, F. Ribeiro, C. Zhang, and D. Florêncio, "L1 regularized room modeling with compact microphone arrays," in *Proc. of ICASSP*, 2010.

[16] S.J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, "An Interior-Point Method for Large-Scale l1-Regularized Least Squares," *IEEE Journal of Selected Topics in Sig. Proc.*, vol. 1, no. 4, pp. 606–617, 2007.

[17] J.M. Bioucas-Dias and M.A.T. Figueiredo, "A new TwIST: two-step iterative shrinkage/thresholding algorithms for image restoration," *IEEE Transactions on Image Processing*, vol. 16, no. 12, pp. 2992–3004, 2007.

[18] M.V. Afonso, J.M. Bioucas-Dias, and M.A.T. Figueiredo, "Fast image recovery using variable splitting and constrained optimization," *IEEE Transactions on Image Processing*, vol. 19, no. 9, pp. 2345–2356, 2010.

[19] T.I. Laakso, V. Välimäki, M. Karjalainen, and U.K. Laine, "Splitting the unit delay—tools for fractional delay filter design," *IEEE Signal Processing Magazine*, vol. 13, no. 1, pp. 30–60, 1996.

[20] S. Muller and P. Massarani, "Transfer-function measurement with sweeps," *J. Audio Eng. Soc.*, vol. 49, no. 6, pp. 443–471, 2001.


**Flávio Ribeiro** received the B.S. degree in electrical engineering from Escola Politécnica, University of São Paulo in 2005, and the B.S. degree in mathematics from the Institute of Mathematics and Statistics, University of São Paulo in 2008. He currently pursuing the Ph.D. degree in electrical engineering, also at Escola Politécnica, University of São Paulo.

From 2007 to 2009, he was a hardware engineer at Licht Labs, where he developed controllers for power transformers and substations. On the summers of 2009, 2010 and 2011, he was a research intern at Microsoft Research Redmond. He won the best student paper award at ICME 2010. His research interests include array signal processing, multimedia signal processing and computational linear algebra.

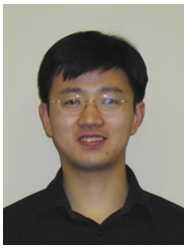
**Dinei Florencio** received the B.S. and M.S. from University of Brasília (Brazil), and the Ph.D. from Georgia Tech, all in Electrical Engineering. He is a researcher with Microsoft Research since 1999, currently with the Multimdia, Interaction, and Communication group. From 1996 to 1999, he was a member of the research staff at the David Sarnoff Research Center. He was also a research co-op student with AT&T Human Interface Lab (now part of NCR) from 1994 to 1996, and a Summer intern at the (now defunct) Interval Research in 1994. Dr. Florencio's current research focus include signal processing and computer security. In the area of signal processing, he works in audio and video processing, with particular focus to real time communication. He has numerous contribution in Speech Enhancement, 3D audio and video, Microphone arrays, Image and video coding, Spectral Analysis, and non-linear algorithms. In the area of computer security, his interest focuses in cybercrime and problems that can be assisted by algorithmic research. Topics include phishing prevention, user authentication, sender authentication, human interactive proofs, and economics of cybercrime.

Dr. Florencio is a senior member of the IEEE, and has published over 50 referred papers, and 36 granted US patents (with another 20 currently pending). He received the 1998 Sarnoff Achievement Award, an NCR inventor award, and a SAIC award. His papers have won awards at SOUPS'2010, ICME'2010, and MMSP'2009. His research has enhanced the lives of millions of people, through high impact technology transfers to many Microsoft products, including Live Messenger, Exchange Server, RoundTable, and the MSN toolbar. He is a member of the ICME and Hot3D steering committees, and an associated editor for the IEEE Trans. on Information Forensics and Security. He is also a member of the ICME IEEE SPS Multimedia Technical Committee, and of the IEEE SPS Information Forensics and Security Technical Committee. Dr. Florencio was general chair of CBSP'2008, MMSP'2009 and WIFS'2011 and technical co-chair of Hot3D'2010, WIFS'2010, ICME'2011, and MMSP'13.

**Demba Ba** graduated from the University of Maryland College Park in May 2004 with the Bachelor of Science degree in Electrical Engineering. He earned the Master of Science and Doctor of Philosophy degrees in Electrical Engineering and Computer Science from the Massachusetts Institute of Technology, respectively in May 2006 and May 2011. In 2006 and 2009, he worked as an summer research intern with the Communication and Collaboration Systems group at Microsoft Research, Redmond. Dr. Ba is currently a postdoctoral associate with the MIT/Harvard Neuroscience Statistics Research Laboratory, where he is developing theory and efficient algorithms to assess synchrony among large assemblies of neurons.

Dr. Ba's current research interests lie in the areas of mathematical and statistical signal processing, with a focus on applications in multimedia and biomedical signal processing, and more generally in domains where the signals of interest have highly non-Gaussian statistics. In this context, Dr. Ba is interested in both developing novel theory, as well as efficient algorithms that scale well to large data sets.

Dr. Ba also has an interest in non-Gaussian adaptive filtering techniques for the task of dynamically updating the ranking of a set of items based on pair wise comparisons. Dr. Ba is exploring applications of these techniques, for example to sports analytics, specifically to ranking of teams involved in a tournament using online, as opposed to batch, win/loss data.



**Cha Zhang** is a Researcher in the Multimedia, Interaction and Communication Group at Microsoft Research (Redmond, WA). He received the B.S. and M.S. degrees from Tsinghua University, Beijing, China in 1998 and 2000, respectively, both in Electronic Engineering, and the Ph.D. degree in Electrical and Computer Engineering from Carnegie Mellon University, in 2004. His current research focuses on applying various machine learning and computer graphics/computer vision techniques to multimedia applications, in particular, multimedia teleconferencing. Dr. Zhang is a Senior Member of IEEE. He was the Publicity Chair for International Packet Video Workshop in 2002, the Program Co-Chair for the first Immersive Telecommunication Conference (IMMERSCOM) in 2007, the Steering Committee Co-Chair and Publicity Chair for IMMERSCOM 2009, the Program Co-Chair for the ACM Workshop on Media Data Integration (in conjunction with ACM Multimedia 2009), and the Poster&Demo Chair for ICME 2011. He served as TPC members for many conferences including ACM Multimedia, CVPR, ICCV, ECCV, MMSP, ICME, ICPR, ICWL, etc. He currently serves as an Associate Editor for Journal of Distance Education Technologies, IPSJ Transactions on Computer Vision and Applications, and ICST Transactions on Immersive Telecommunications. He was a guest editor for Advances in Multimedia, Special Issue on Multimedia Immersive Technologies and Networking.