3D Room Geometry Inference Using a Linear Loudspeaker Array and a Single Microphone

Cagdas Tuna, Antonio Canclini, Federico Borra, Philipp Götz, Fabio Antonacci, *Member, IEEE*, Andreas Walther, Augusto Sarti, *Senior Member, IEEE*, and Emanuël A. P. Habets, *Senior Member, IEEE*

***Abstract*— Sound reproduction systems may highly benefit from detailed knowledge of the acoustic space to enhance the spatial sound experience. This article presents a room geometry inference method based on identification of reflective boundaries using a high-resolution direction-of-arrival map produced via room impulse responses (RIRs) measured with a linear loudspeaker array and a single microphone. Exploiting the sparse nature of the early part of the RIRs, Elastic Net regularization is applied to obtain a 2D polar-coordinate map, on which the direct path and early reflections appear as distinct peaks, described by their propagation distance and direction of arrival. Assuming a separable room geometry with four side-walls perpendicular to the floor and ceiling, and imposing pre-defined geometrical constraints on the walls, the 2D-map is segmented into six regions, each corresponding to a particular wall. The salient peaks within each region are selected as candidates for the first-order wall reflections, and a set of potential room geometries is formed by considering all possible combinations of the associated peaks. The room geometry is then inferred using a cost function evaluated on the higher-order reflections computed via beam tracing. The proposed method is tested with both simulated and measured data.**

***Index Terms*—DOA estimation, reflector localization, echo labeling, room geometry inference.**

I. Introduction

INFORMATION on the acoustic environment is of importance in advanced audio systems, improving the system performance and enabling new functionalities in applications. For instance, the knowledge of the room geometry can be used to increase robustness in sound source localization [1], [2], enhance the target signal in dereverberation [3], [4], and improve the spatial impression through room compensation in sound reproduction [5]. Room geometry inference (RGI) methods generally involve the localization of first-order reflections from the walls enclosing the room via the room impulse responses (RIRs) measured between different arrangements of loudspeakers and microphones. RIRs are processed to extract times of arrivals (TOAs) of the direct path and reflections, which are then used to estimate the locations and orientations of wall reflectors.

Reflector localization requires TOAs to be grouped into sets, each associated with a particular reflective boundary, which is known as TOA disambiguation or echo labeling in the literature [6], [7]. A class of reflector localization methods uses an ellipse-based formulation [8]–[14], where a reflector is identified by finding the common tangent to all the ellipses with the foci given by the pair of speaker- and microphone-positions traced from the corresponding TOAs. A 2D reflector localization algorithm to obtain the common tangent is proposed in [8], and its robustness to temperature variations is investigated in [9]. This approach is also adopted in [10] to blindly localize the walls in 2D using single-source multiple-receiver setups. In [11], the use of a compact microphone array guarantees that TOAs belonging to the same reflector in 2D are ordered consistently in different RIRs from the same array, except in some special cases. An extension to 3D environments is achieved in [12] by decomposing the 3D microphone array into 2D sub-arrays and in [13], [14] by using ellipsoids. In an alternative reflector localization method proposed in [15], the 3D room geometry is inferred from the knowledge of TOAs of first-order reflections obtained via the Euclidean distance matrix associated with a distributed microphone array. In [16], TOAs are acquired from an uncalibrated 3D setup by means of an auto-localization algorithm proposed in [17], and the room geometry is then estimated using a greedy iterative approach.

Another class of reflector localization techniques bypasses the need for TOA detection in each RIR by jointly processing the array measurements to generate polar time-domain TOA/DOA (direction-of-arrival) maps describing the evolution of planar reflections following the direct path over time, which emerge on these maps as distinct peaks [18]–[22]. In [18], the TOA-space maps are generated using plane-wave decomposition in frequency domain over RIRs measured with a circular array. In [19], [20], the superdirective array (SDA) beamformer [23] is applied to RIRs recorded with bi-circular arrays to obtain time-domain azimuth-DOA maps to be used for 3D reflector localization and classification. Exploiting the acoustic reciprocity principle [24], the Linear Radon Transform (LRT) is applied to the stack of RIRs measured with a linear loudspeaker array and a

Manuscript received July 16, 2019; revised January 8, 2020 and April 24, 2020; accepted May 18, 2020. Date of publication May 28, 2020; date of current version November 7, 2020. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Maria de Diego. *(Corresponding author: Cagdas Tuna.)*

Cagdas Tuna and Emanuël A. P. Habets are with the International Audio Laboratories Erlangen (A joint institution of the Friedrich-Alexander-University Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits IIS), 91058 Erlangen, Germany (e-mail: cagdas.tuna@iis.fraunhofer.de; emanuel.habets@audiolabs-erlangen.de).

Antonio Canclini, Federico Borra, Fabio Antonacci, and Augusto Sarti are with the Dipartimento di Elettronica Informazione e Bioingegneria, Politecnico di Milano, 20133 Milan, Italy (e-mail: antonio.canclini@polimi.it; federico.borra@polimi.it; fabio.antonacci@polimi.it; augusto.sarti@polimi.it).

Philipp Götz and Andreas Walther are with the Fraunhofer Institute for Integrated Circuits IIS, 91058 Erlangen, Germany (e-mail: philipp.goetz@iis.fraunhofer.de; andreas.walther@iis.fraunhofer.de).

Digital Object Identifier 10.1109/TASLP.2020.2998299

single microphone in [21] to generate similar maps for 2D TOA disambiguation.

Most existing reflector localization methods require the use of a single omnidirectional speaker and a microphone array that has the same dimension as the case of interest (e.g., circular or planar arrays for 2D reflector localization and spherical arrays for the 3D case). If these techniques are implemented using setups with lower-dimensional arrays, they may fail to accurately localize reflectors and produce probable but incorrect estimates due to the geometrical ambiguity arising from the lack of a dimension in the array. There are few techniques in the literature circumventing this geometrical ambiguity introduced by the transducer array. In [25], a compact circular array is used for 3D RGI, and a l_1-regularized least-squares fit is applied on the measured RIRs using an extensive set of acoustic reflections synthetically generated in 3D. In [14], [20], bi-circular microphone arrays are used for 3D reflection localization, and the up/down ambiguity is circumvented by assuming that the array positioning relative to the floor/ceiling is known *a priori* (e.g., floor is closer than the ceiling). As a follow-up to the LRT-based technique developed in [21], a measurement setup composed of a 2D rectangular loudspeaker array (with the intention to place it around a video screen) and an omnidirectional microphone is used in [22] for 3D RGI and the front/back ambiguity is avoided by assuming the loudspeaker array is placed close to one of the walls at a nearly-parallel position.

In this manuscript, we introduce a 3D RGI algorithm using a setup consisting only of a linear loudspeaker array and a single omnidirectional microphone. With the increasing use of soundbars in TV sets, 3D reflector localization with a linear loudspeaker array becomes more relevant to commercial applications. The performance of the sound reproduction system can significantly be improved if it is made aware of the surrounding environment (e.g., [26]), as some prior knowledge of the room geometry such as the floor-map and floor/ceiling height is key to sound rendering in achieving a pleasing immersive experience. However, to the best of our knowledge, 3D RGI with a linear array (1D configuration), where the spatial diversity is significantly reduced in two dimensions, is yet to be addressed in the literature. As the first attempt to combat this challenging situation, the main contributions of this work include:

1) a novel sparsity-constrained high-resolution 2D polar-space DOA mapping technique using RIRs recorded with a synchronized setup made up of a linear loudspeaker array and a single microphone,
2) a semisupervised approach tackling the geometrical ambiguity and identifying potential first-order reflection candidates through the segmentation of the DOA map into bounded regions, each associated with a wall, based on pre-defined constraints for wall dimensions and orientations,
3) room geometry inference based on a cost function measuring the match between the higher-order reflections estimated via beam tracing [27]–[29] and the actual reflections spotted on the DOA map.

The proposed 3D RGI algorithm is validated with an extensive set of RIRs measured in rooms with different wall characteristics and reverberation times as well as a simulated replica of one of these rooms.

The remainder of the manuscript is organized as follows: Section II gives an overview of the proposed method. Section III provides the description of a separable room geometry. Section IV formulates the DOA-map estimation from RIR measurements, and explains the peak detection and pruning method. Section V details the segmentation of the DOA map into regions generated based on the geometric constraints imposed on the walls. Section VI explains how the inferred room geometry is determined via beam tracing from all possible room configurations. Section VII presents the performance evaluation of the proposed method with real and simulated data and Section VIII concludes the manuscript.

II. Method Overview

The proposed 3D RGI algorithm using a linear loudspeaker array and a single microphone consists of three main steps, as outlined in Fig. 1.

To tackle the geometrical ambiguity arising from the limited spatial diversity of the linear array, a room of interest is assumed to have a separable geometry (i.e., side-walls are perpendicular to floor and ceiling and no discontinuity exists along the walls). In addition, the microphone is assumed to be in line of sight with the loudspeakers located in front of the array on the same horizontal plane, parallel to the floor.

In the first step of the proposed RGI algorithm, RIRs recorded synchronously between the linear loudspeaker array and single microphone (i.e., any initial delay in a given RIR is associated with the acoustic propagation path) are used to generate a DOA map, where the localization of the real- and image-microphones in 3D is achieved by their projection onto the 2D polar-coordinate space. One should expect the resulting DOA map to have a small number of discrete peaks resulting from the sparse nature of the early parts of RIRs containing only the direct path and distinct early reflections. A high-resolution map is thus obtained by formulating DOA estimation as a linear inverse problem and then solving it via Elastic Net regularization [30], [31], in which a penalty term composed of a weighted summation of LASSO and ridge-regression constraints is used in the optimization cost function. A standard 2D peak detection algorithm [21] followed by a peak pruning procedure is then utilized on the estimated DOA map to identify the salient peaks likely to be associated with the direct path and early reflections.

In the second step, the DOA map is segmented into multiple bounded regions, each corresponding to a wall, to determine the peak candidates that may potentially correspond to the first-order reflections. These bounded regions are generated based on a set of relaxed geometrical constraints imposed on the walls to restrict the search space for the reflecting boundary surfaces. In more detail, lower and upper bounds are pre-defined for the distances from the loudspeaker array to the walls along with an angular limit on the orientation of side-walls, allowing some degree of deviation from a shoe-box room model. Within each region, the most significant peaks are then selected as potential wall candidates.

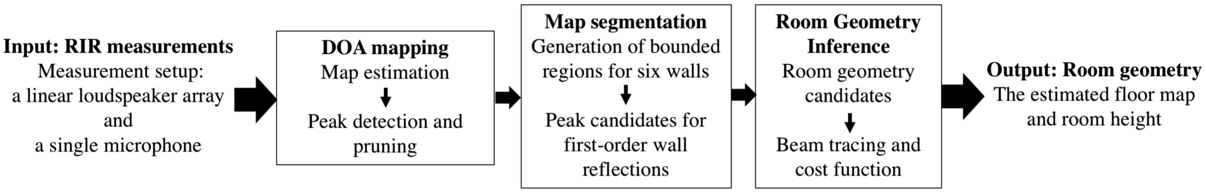
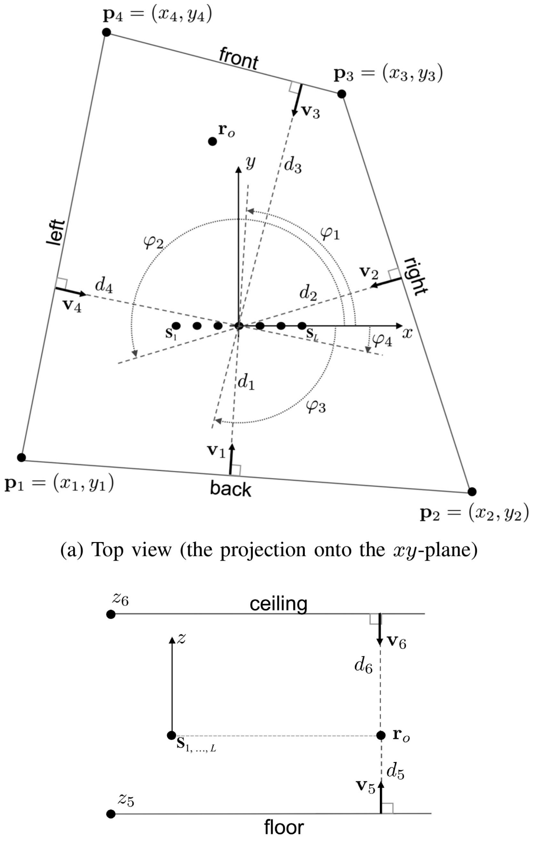


Fig. 1. The block diagram of the proposed 3D room geometry inference algorithm using a linear loudspeaker array and a single microphone.

In the third step, all possible room geometries are obtained through a Cartesian product of the sets of wall candidates extracted from the six bounded regions that correspond to four side-walls, floor and ceiling. Finally, the room geometry described by its floor map and height is inferred by using a cost function that evaluates the agreement between the higher-order reflections estimated from the candidate first-order reflections via the beam-tracing method [27]–[29] and the peaks on the actual DOA map.



(a) Top view (the projection onto the xy-plane)

(b) Side view (the projection onto the yz-plane)

Fig. 2. The separable room geometry obtained by the Cartesian product of 2D and 1D environments (2D × 1D): The room shape can be fully described by its floor map (top view) and height (side view).

III. Separable Room Geometry

A. Room Description

Under the assumption of a separable room geometry, which is obtained by the Cartesian product of 2D and 1D environments (2D × 1D), a room of an unknown shape can be completely described by its floor map and height, having perfectly flat side-walls of equal height and all perpendicular to the floor and ceiling. The setup used in this manuscript is illustrated in Fig. 2. The inference of a 3D room geometry is achieved by estimating the parameter pair $(\mathbf{v}_w, d_w)$, the normal vector and distance from the origin, for each of the six walls $w = 1, \ldots, 6$ via RIRs measured from a linear array of loudspeakers located parallel to the floor and ceiling at the positions $\mathbf{s}_l$, $\forall l \in \{1, \ldots, L\}$ to an omnidirectional microphone located at $\mathbf{r}_o$. Without loss of generality, a reference system with the origin located at the array center and the x-axis aligned with the array is considered such that the lth array element is positioned at $\mathbf{s}_l = [x_{\mathbf{s}_l}, 0, 0]^T$. The microphone is positioned to be in front of the speaker array at the same height (i.e., $\mathbf{r}_o = [x_o, y_o, z_o]^T$, where $y_o > 0$ and $z_o = 0$). Under the 2D × 1D geometry assumption, each side-wall may be described by the orientation angle φ_w yielding the normal vector $\mathbf{v}_w = [\cos\varphi_w, \sin\varphi_w, 0]^T$, and the distance d_w $(w = 1, \ldots, 4)$, whereas the normal vectors are known *a priori* for floor and ceiling given by $\mathbf{v}_5 = [0, 0, 1]^T$ and $\mathbf{v}_6 = [0, 0, -1]^T$ such that only the corresponding distances d_5 and d_6 need to be estimated. Under the aforementioned assumptions, it readily follows that the room shape is determined by a total of 10 parameters (4 angles and 6 distances).

A set of geometric constraints is imposed on the walls to limit the search space for the room geometry to be inferred. For side-walls, the search for the orientation angle is restricted into the range $\varphi_w \in [\overline{\varphi}_w - \Delta\varphi_w, \overline{\varphi}_w + \Delta\varphi_w]$, where $\Delta\varphi_w$ denotes the maximum allowed deviation from a shoe-box room model, in which the orientation angles $\overline{\varphi}_w$, $w = 1, \ldots, 4$ are specified for four side-walls as $\overline{\varphi}_1 = \pi/2$ (back wall), $\overline{\varphi}_2 = \pi$ (right wall), $\overline{\varphi}_3 = -\pi/2$ (front wall) and $\overline{\varphi}_4 = 0$ (left wall). This leads to the angular constraint for side-walls $(w = 1, \ldots, 4)$

$$\arccos\langle \mathbf{v}_w, \mathbf{v}_{\overline{\varphi}_w} \rangle \leq \Delta\varphi_w, \tag{1}$$

where $\mathbf{v}_{\overline{\varphi}_w}$ refers to the normal vector of the wth wall in a shoe-box room and $\langle \cdot, \cdot \rangle$ denotes the inner product. For each of the six walls, the distance d_w is bounded by pre-defined minimum and maximum values $(w = 1, \ldots, 6)$:

$$d_w^{\min} \leq d_w \leq d_w^{\max}. \tag{2}$$

B. Geometrical Ambiguity

Assuming a linear array setup with the origin located at the array center and the x-axis aligned with the array, each point $\mathbf{r}$ in 3D-space may be projected onto the 2D polar-coordinate space

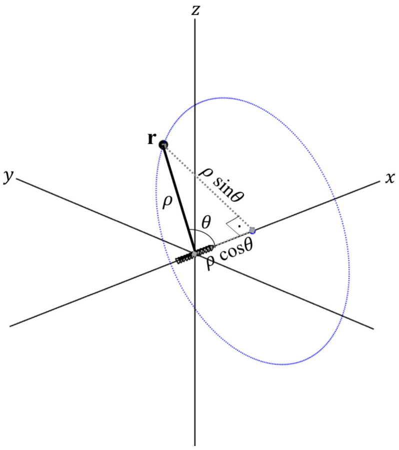


Fig. 3. Geometrical ambiguity due to using a linear array: The pair (ρ,θ) may correspond to any point $\mathbf{r}$ located on the circumference of a circle perpendicular to the xy-plane, centered at $[\rho\cos\theta,0,0]^T$ with the radius equal to $\rho\sin\theta$, resulting in a circular symmetry for the linear array geometry.

with its polar coordinates given by the pair (ρ,θ):

$$\rho = \|\mathbf{r}\| \qquad \text{and} \qquad \theta = \arccos\frac{\langle \mathbf{r},\mathbf{u}\rangle}{\|\mathbf{r}\|}\ , \tag{3}$$

where ρ is the propagation distance, the angle $\theta \in [0°, 180°]$ and $\mathbf{u} = [1,0,0]^T$ is the unit vector in the direction of x axis. The mapping from $\mathbf{r}$ to the polar-coordinate space can be unequivocally determined, but the reverse is not true, since this is not an invertible operation. As illustrated in Fig. 3, the pair (ρ,θ) may represent any point on the circumference of a circle perpendicular to the xy-plane, centered at $[\rho\cos\theta,0,0]^T$ with the radius equal to $\rho\sin\theta$, indicating a circular symmetry for the linear array geometry when used in 3D setups.

The geometrical ambiguity arising from the linear array setup is lessened under the assumption of a 2D × 1D separable room geometry. Side-walls generate image-microphone positions all lying on the same plane with the speaker array and measurement position (i.e., the xy-plane) such that the uncertainty reduces down to a "front-back" ambiguity, where it is still not possible to distinguish between the image-microphones occurring in the "front" ($y > 0$) or "back" ($y < 0$) of the speaker array. Floor and ceiling can be treated separately from side-walls because the first-order image-microphones (as well as higher-order reflections between floor and ceiling) should be then located in 1D space (i.e., the line parallel to the z-axis and passing from $\mathbf{r}_o$), resulting in an "up" ($z > 0$) and "down" ($z < 0$) ambiguity for floor and ceiling.

Since the measurement point is also assumed to be positioned in front of the speaker array at the same height, the Cartesian coordinates of $\mathbf{r}_o$ are given by the corresponding the polar-coordinate pair (ρ_o,θ_o):

$$\mathbf{r}_o = \begin{bmatrix} \rho_o\cos\theta_o \\ \rho_o\sin\theta_o \\ 0 \end{bmatrix}. \tag{4}$$

Based on geometrical acoustics and the reciprocity principle, the specular reflection from a wall described by the parameter pair $(\mathbf{v},d)$ can be regarded as an acoustic path originated at the first-order image microphone position $\mathbf{r}'_o$, which is obtained by mirroring $\mathbf{r}_o$:

$$\mathbf{r}'_o = (\mathbf{I} - 2\mathbf{v}\mathbf{v}^T)\mathbf{r}_o - 2d\mathbf{v}. \tag{5}$$

Given the $\mathbf{r}_o$ and $\mathbf{r}'_o$, the wall parameters may then be fully determined by using (5). The normal vector is given by

$$\mathbf{v} = \frac{\mathbf{r}_o - \mathbf{r}'_o}{\|\mathbf{r}_o - \mathbf{r}'_o\|}, \tag{6}$$

and the distance is computed via

$$d = -\frac{1}{2}\mathbf{v}^T(\mathbf{r}_o + \mathbf{r}'_o). \tag{7}$$

1) Side-Walls: Due to the front-back ambiguity, a pair (ρ,θ) in polar-coordinate space representing a first-order side-wall reflection may be associated with two distinct positions in geometric space, which are given by

$$\mathbf{r}'_{o,y^-} = \begin{bmatrix} \rho\cos\theta \\ -\rho\sin\theta \\ 0 \end{bmatrix} \quad \text{and} \quad \mathbf{r}'_{o,y^+} = \begin{bmatrix} \rho\cos\theta \\ \rho\sin\theta \\ 0 \end{bmatrix}, \tag{8}$$

Thus, the mapping between polar-coordinate space and side-wall parameters can be summarized as

$$(\rho,\theta) \overset{\mathbf{r}'_{o,y^-}}{\longleftrightarrow} (\mathbf{v}_{y^-}, d_{y^-}),$$

$$(\rho,\theta) \overset{\mathbf{r}'_{o,y^+}}{\longleftrightarrow} (\mathbf{v}_{y^+}, d_{y^+}), \tag{9}$$

where the wall-parameter pairs $(\mathbf{v}_{y^-}, d_{y^-})$ and $(\mathbf{v}_{y^+}, d_{y^+})$ are computed via (6) and (7).

2) Floor and Ceiling: Under the separable geometry assumption, the image-microphones corresponding to the first-order floor and ceiling reflections share the same x-y coordinates with the measurement point $\mathbf{r}_o$ with an up-down ambiguity. The two image-microphone positions are thus given by

$$\mathbf{r}'_{o,z^-} = \begin{bmatrix} \rho_o\cos\theta_o \\ \rho_o\sin\theta_o \\ -\sqrt{\rho^2-\rho_o^2} \end{bmatrix} \quad \text{and} \quad \mathbf{r}'_{o,z^+} = \begin{bmatrix} \rho_o\cos\theta_o \\ \rho_o\sin\theta_o \\ \sqrt{\rho^2-\rho_o^2} \end{bmatrix}. \tag{10}$$

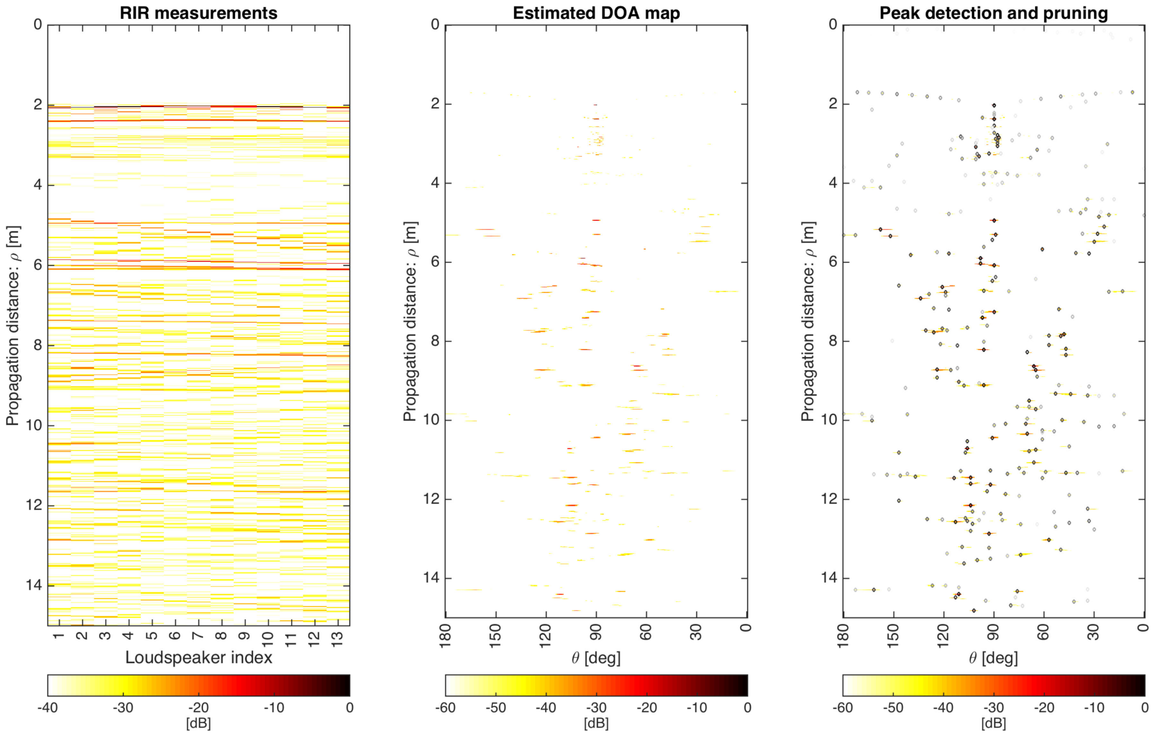
where the sign ambiguity is resolved, as $\mathbf{r}'_{o,z^-}$ must correspond to a floor reflection and $\mathbf{r}'_{o,z^+}$ to a ceiling reflection. Given the normal vectors, $\mathbf{v}_5 = [0,0,1]^T$ and $\mathbf{v}_6 = [0,0,-1]^T$ for floor and ceiling, the distance parameters computed via (7) attain the same value for both $\mathbf{r}'_{o,z^-}$ and $\mathbf{r}'_{o,z^+}$:

$$d_z = \frac{\sqrt{\rho^2-\rho_o^2}}{2}. \tag{11}$$

This results in the mapping between polar-coordinate space and floor/ceiling wall parameters as

$$(\rho,\theta) \overset{\mathbf{r}'_{o,z^-}}{\longleftrightarrow} (\mathbf{v}_5, d_z),$$

$$(\rho,\theta) \overset{\mathbf{r}'_{o,z^+}}{\longleftrightarrow} (\mathbf{v}_6, d_z). \tag{12}$$


RIR measurements

Estimated DOA map

Peak detection and pruning

Fig. 4. DOA mapping based on RIRs recorded between a uniform loudspeaker array and a single microphone: The measured RIRs (illustrated here as a normalized and dB-scaled image of vertically stacked RIRs) are first translated into a 2D polar-coordinate DOA map (shown as a normalized and dB-scaled image) and the salient peaks that are likely to correspond to real- and image-microphones are then identified on the DOA map using a standard 2D peak detecting and pruning algorithm (peaks are colored in gray scale with respect to their magnitude).

IV. DOA Mapping

An illustrative summary of the DOA-mapping procedure described in this section is given in Fig. 4. To combat the geometrical ambiguity resulting from the lack of spatial diversity and identify the locations and orientations of wall reflectors, RIRs recorded between the linear loudspeaker array and single microphone are first translated into a 2D polar-coordinate map, onto which real- and image-microphones in 3D are projected, emerging as distinct peaks described by their propagation distances and DOAs. A peak detection and pruning procedure is then applied to the estimated map to determine the significant peaks that are likely to correspond to real- and image-microphones.

A. Forward Problem

Let $X(\rho,\theta)$ be the unknown 2D discrete polar-coordinate map of size $M \times I$, where the 2D-grid has a length of M along the ρ-axis and a length of I along the θ-axis. Using the acoustic reciprocity principle [24], a grid point located at (ρ_m,θ_i) on the map corresponds to a 2D projection of a potential point source associated with a real- or image-microphone position in 3D. Assuming all acoustic devices are synchronized and taking into account all the points across the 2D-grid, the nth element of the discrete-time room impulse response vector $\mathbf{h}^{(l)} \in \mathbb{R}^N$ of length N acquired between the speaker l in the linear array and the single microphone may be expressed as

$$h^{(l)}(n) = \sum_{m=0}^{M-1}\sum_{i=0}^{I-1} h^{(l)}_{\rho_m,\theta_i}(n) + b^{(l)}(n), \qquad (13)$$

where $b^{(l)}(n)$ is the nth element of the additive measurement noise vector $\mathbf{b}^{(l)} \in \mathbb{R}^N$. Considering a spherical wave propagation model, $h^{(l)}_{\rho_m,\theta_i}(n)$ described by the contribution of a particular grid point located at (ρ_m,θ_i) to $h^{(l)}(n)$ is given by

$$h^{(l)}_{\rho_m,\theta_i}(n) \approx X_{\rho_m,\theta_i}\frac{\delta(n-\lfloor f_s \cdot d^{(l)}_{\rho_m,\theta_i}/c\rceil)}{d^{(l)}_{\rho_m,\theta_i}}, \qquad (14)$$

where $d^{(l)}_{\rho_m,\theta_i} = [(\rho_m\cos\theta_i - x_{\mathbf{s}_l})^2 + (\rho_m\sin\theta_i)^2]^{1/2}$ is the propagation distance from the point (ρ_m,θ_i) to the speaker l with the corresponding time-of-flight $d^{(l)}_{\rho_m,\theta_i}/c$ (c: the speed of sound), $\delta(n)$ is the discrete-time Dirac delta function, $\lfloor . \rceil$ denotes the operator for rounding to the closest integer value and f_s is the sampling frequency. The amplitude of the grid point denoted by X_{ρ_m,θ_i} only takes nonzero values if a real- or image-microphone is actually projected onto the point (ρ_m,θ_i), representing a numerical indicator for a combination of potential acoustic phenomena including the level of absorption at a reflective surface, diffraction, and the microphone and loudspeaker directivity.

The rounding operation in (14) results in a mismatch between the actual time delay and sampled instance, and hence, may cause some noticeable distortion due to the high sensitivity to the relative phase between the speakers in the array. Furthermore, in practice, the measured RIRs rarely contain sharp peaks as a result of the factors including the type of excitation signal and equipment effects, etc. Alternatively, the Dirac delta function may be replaced by a function, which models these effects by "smoothing" the relation between the grid points on the map $X(\rho,\theta)$ and the discrete-time RIRs [32], [33]. Let $\mathbf{g}^{(l)}_{\rho_m,\theta_i}$ be a 1D smoothing function of length N sampled from a continuous-time function $g(t)$ shifted by the actual time-delay $d^{(l)}_{\rho_m,\theta_i}/c$, whose nth element is given by

$$g^{(l)}_{\rho_m,\theta_i}(n) = g(t - d^{(l)}_{\rho_m,\theta_i}/c)\Big|_{t=\frac{n}{f_s}}, \quad 0 \leq n \leq N-1. \tag{15}$$

The revised $h^{(l)}(n)$ can then be given by

$$h^{(l)}(n) = \sum_{m=0}^{M-1}\sum_{i=0}^{I-1} X_{\rho_m,\theta_i}\frac{g^{(l)}_{\rho_m,\theta_i}(n)}{d^{(l)}_{\rho_m,\theta_i}} + b^{(l)}(n). \tag{16}$$

Let $\mathbf{A}^{(l)}_{\rho_m} = [\mathbf{g}^{(l)}_{\rho_m,\theta_0}/d^{(l)}_{\rho_m,\theta_0}, \ldots, \mathbf{g}^{(l)}_{\rho_m,\theta_{I-1}}/d^{(l)}_{\rho_m,\theta_{I-1}}]$ be the angular-grid matrix at the propagation distance ρ_m for the lth array element and $\mathbf{x}_{\rho_m} = [X_{\rho_m,\theta_0}, \ldots X_{\rho_m,\theta_{I-1}}]^T$ be the amplitude vector across the DOAs at the propagation distance ρ_m corresponding to the mth column of the 2D-map $X(\rho,\theta)$. Then, vertically concatenating the impulse responses from L speakers and reformulating them in the matrix form yields the additive signal-noise model for the linear inverse problem:

$$\underbrace{\begin{bmatrix}\mathbf{h}^{(1)}\\ \vdots \\ \mathbf{h}^{(L)}\end{bmatrix}}_{\boldsymbol{h}} = \underbrace{\begin{bmatrix}\mathbf{A}^{(1)}_{\rho_0} & \cdots & \mathbf{A}^{(1)}_{\rho_{M-1}}\\ \vdots & \ddots & \vdots \\ \mathbf{A}^{(L)}_{\rho_0} & \cdots & \mathbf{A}^{(L)}_{\rho_{M-1}}\end{bmatrix}}_{\mathbf{A}}\underbrace{\begin{bmatrix}\mathbf{x}_{\rho_0}\\ \vdots \\ \mathbf{x}_{\rho_{M-1}}\end{bmatrix}}_{\mathbf{x}} + \underbrace{\begin{bmatrix}\mathbf{b}^{(1)}\\ \vdots \\ \mathbf{b}^{(L)}\end{bmatrix}}_{\mathbf{b}}, \tag{17}$$

where $\mathbf{A} \in \mathbb{R}^{L\cdot N \times M\cdot I}$ is the concatenated steering matrix, the vector $\mathbf{x} \in \mathbb{R}^{M\cdot I}$ corresponds to the polar-coordinate map $X(\rho,\theta)$ in the vectorized form, and $\mathbf{h} \in \mathbb{R}^{L\cdot N}$ and $\mathbf{b} \in \mathbb{R}^{L\cdot N}$ are the vertically-concatenated impulse-response and additive measurement-noise vectors, respectively.

B. Map Estimation

The early parts of the measured RIRs consist of the direct path and acoustic reflections from the room surfaces, which in turn should appear on the DOA map as a relatively small number of distinct peaks, prompting the use of sparse signal estimation techniques. The Elastic Net regularization is considered here, with its estimate given by [30], [31]

$$\begin{aligned}\hat{\mathbf{x}} = \underset{\mathbf{x}\geq 0}{\operatorname{argmin}}\, &\frac{1}{2}||\mathbf{h} - \mathbf{A}\mathbf{x}||_2^2 \\ &+ \lambda\left(\beta||\mathbf{x}||_1 + \frac{(1-\beta)}{2}||\mathbf{x}||_2^2\right),\end{aligned} \tag{18}$$

where $\lambda > 0$ is the regularization parameter for the Elastic Net penalty, which is composed of a weighted summation of l_1-norm (LASSO) and l_2-norm (ridge regression) constraints controlled by the weighting parameter $0 \leq \beta \leq 1$. A non-negativity constraint is additionally imposed on $\mathbf{x}$, taking only the interactions between the air and hard surfaces into account. As a result of the design of the matrix $\mathbf{A} \in \mathbb{R}^{L\cdot N \times M\cdot I}$ using a smoothing function in (15), the grid points neighboring a peak on the DOA map should be highly correlated, generating a locally-smooth region in the proximity of the peak. However, one of the known limitations of using only a LASSO penalty is that in underdetermined systems (i.e., $L \cdot N < M \cdot I$), LASSO may only select one variable from a group of highly correlated variables and discard the rest, which may be prevented by including an additional quadratic penalty term in the optimization problem [31]. Therefore, the Elastic Net regularization can achieve a desired level of sparsity while simultaneously selecting the grouped variables by jointly imposing the LASSO and ridge-regression penalties [30], [31].

The Elastic Net may be solved as a LASSO problem using an augmented forward model as follows [31]:

$$\tilde{\mathbf{h}} = \begin{bmatrix}\mathbf{h}\\ \mathbf{0}\end{bmatrix} \quad \text{and} \quad \tilde{\mathbf{A}} = \begin{bmatrix}\mathbf{A}\\ \sqrt{\lambda\cdot(1-\beta)}\,\mathbf{I}\end{bmatrix}. \tag{19}$$

Rewriting (18) yields the LASSO estimate

$$\hat{\mathbf{x}} = \underset{\mathbf{x}\geq 0}{\operatorname{argmin}}\, \frac{1}{2}||\tilde{\mathbf{h}} - \tilde{\mathbf{A}}\mathbf{x}||_2^2 + (\lambda\cdot\beta)||\mathbf{x}||_1. \tag{20}$$

The coordinate descent algorithm is used here to obtain the LASSO solution, during which $\hat{\mathbf{x}}$ is updated cyclically at each iteration [34]. Exploiting the sparse nature of the measurement matrix $\tilde{\mathbf{A}}$, the LASSO estimate may be computed efficiently, since the coordinate descent solution does not require any matrix inversion. Initializing the algorithm with $\hat{\mathbf{x}}^{(-1)} = \mathbf{0}$, the cyclic update at iteration $k \geq 0$ is achieved by minimizing only over $x(\eta)$, the ηth element of $\mathbf{x}$, while keeping all $x(\psi)$, $\psi \neq \eta$ fixed, and the non-negativity constraint is satisfied by zeroing any negative estimate occurred after an update. This leads to a one-sided soft-thresholding operation, which may be expressed as

$$\hat{x}(\eta) = \begin{cases}(\tilde{\mathbf{a}}_\eta^T\tilde{\mathbf{r}}_\eta - (\lambda\cdot\beta))/||\tilde{\mathbf{a}}_\eta||_2^2 & \text{if } \tilde{\mathbf{a}}_\eta^T\tilde{\mathbf{r}}_\eta > (\lambda\cdot\beta)\\ 0 & \text{otherwise}\end{cases}, \tag{21}$$

where the residual vector is given by $\tilde{\mathbf{r}}_\eta = \tilde{\mathbf{h}} - \sum_{\psi\neq\eta}\hat{x}(\psi)\tilde{\mathbf{a}}_\psi$, and $\tilde{\mathbf{a}}_\eta$ and $\tilde{\mathbf{a}}_\psi$ denote the ηth and ψth ($\psi \neq \eta$) columns of $\tilde{\mathbf{A}}$, respectively. The coordinate descent algorithm is stopped when the relative error at iteration k reaches below a pre-defined threshold: $\epsilon^{(k)} < \tau$, where

$$\epsilon^{(k)} = \frac{||\hat{\mathbf{x}}^{(k)} - \hat{\mathbf{x}}^{(k-1)}||_1}{||\hat{\mathbf{x}}^{(k-1)}||_1}. \tag{22}$$

C. Peak Detection and Pruning

The salient peaks on the DOA map $X(\rho,\theta)$ are detected based on a standard 2D peak-picking algorithm to extract the local maxima from an image representation of a 2D function (c.f., [21]). To reduce the set of all detected peaks into a subset of the most significant peaks that are more likely to be associated with actual real- and image-microphones, a pruning procedure is additionally applied, during which all secondary peaks in the neighborhood of more significant ones are discarded in an iterative fashion.

Let $\mathcal{Q}_0$ be the set of all extracted peaks, which are then sorted in descending order with respect to the magnitude and

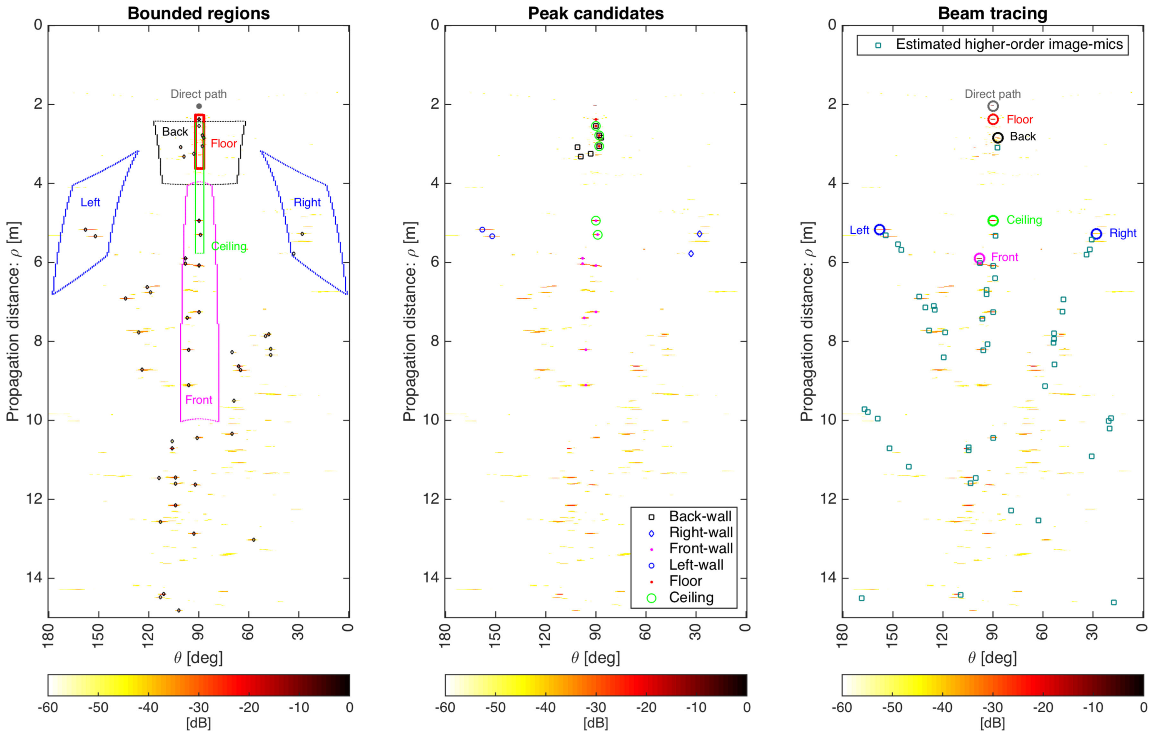


Fig. 5. Map segmentation and room geometry inference via beam tracing: The DOA map is segmented into six regions, each corresponding to a wall, based on the pre-defined constraints on the wall dimensions and orientations, followed by the selection of peaks inside each region that are potential first-order reflection candidates. The room geometry is inferred by using a cost function evaluating the agreement between the higher-order image-microphones estimated from the candidate first-order reflections via beam tracing and the peaks on the actual DOA map.

$\overline{\mathcal{Q}}_{-1} = \emptyset$ be the set of most significant peaks, which is empty at initialization. At iteration $k \geq 0$, all the peaks within the neighborhood of the first element of the set $\mathcal{Q}_k$ denoted by $\overline{\mathbf{q}}_k = (\overline{\rho}_k, \overline{\theta}_k)$ are pruned:

$$\mathcal{Q}_{k+1} = \mathcal{Q}_k \setminus \mathcal{N}(\overline{\mathbf{q}}_k), \tag{23}$$

where $\mathcal{N}(\overline{\mathbf{q}}_k)$ is the rectangular neighborhood of $\overline{\mathbf{q}}_k$, whose size is specified by the thresholds $|\rho - \overline{\rho}_k| \leq T_\rho$ and $|\theta - \overline{\theta}_k| \leq T_\theta$. Then, $\overline{\mathbf{q}}_k$ is added to the set of most significant peaks:

$$\overline{\mathcal{Q}}_k = \overline{\mathcal{Q}}_{k-1} \cup \overline{\mathbf{q}}_k. \tag{24}$$

This procedure is repeated until $\mathcal{Q}_{k+1} = \emptyset$. The set of the most significant peaks is finally obtained as

$$\overline{\mathcal{Q}} = \{\overline{\mathbf{q}}_0, \overline{\mathbf{q}}_1, \ldots\}. \tag{25}$$

V. Map Segmentation

Fig. 5 shows an example of the generation of bounded regions using the geometric constraints defined in (1) and (2), each corresponding to either a side-wall or floor/ceiling, followed by the selection of the peaks inside each region as candidates for first-order reflections. Such a map-segmentation procedure involves the remapping of the salient peaks given by their propagation-path distance and DOA-angle in the 2D polar-coordinate space back onto the 3D geometric space to identify the real- and image-microphone positions corresponding to the direct path and first-order wall reflections for room geometry inference.

A. Direct Path Estimation

Assuming that the measurement point $\mathbf{r}_o$ is in line of sight with all array elements, $\mathbf{s}_l$, $\forall l \in \{1, \ldots, L\}$, the first arrival corresponding to the direct path should generally emerge on the DOA map as the peak with the highest magnitude. However, in some cases, there may also be peaks with amplitudes higher than the direct path as a result of compensating for the propagation-distance attenuation based on the spherical-wave propagation model. Taking a practical approach, the peak corresponding to the direct path is searched among the ten largest peaks in $\overline{\mathcal{Q}}$ and the first arrival among them is estimated as the peak corresponding to the direct path, which is labelled as the pair (ρ_o, θ_o).

B. Generation of Bounded Regions

1) Side-Walls: A pair (ρ, θ) becomes a valid candidate for the wth side-wall provided that the conditions described by (1) and (2) are satisfied by $(\mathbf{v}_{y^+}, d_{y^+})$ and/or $(\mathbf{v}_{y^-}, d_{y^-})$ given in (9). The set of all valid pairs (ρ, θ) then forms two bounded regions $\mathcal{M}_{y^\pm}^{(w)}$ on the DOA map for the side-walls $w = 1, \ldots, 4$, which may be formally defined as

$$\mathcal{M}_{y^\pm}^{(w)} = \left\{ (\rho, \theta) \ : \begin{array}{c} \arccos \langle \mathbf{v}_{y^\pm}, \mathbf{v}_{\bar{\alpha}_w} \rangle \leq \Delta\varphi_w \\ d_w^{\min} \leq d_{y^\pm} \leq d_w^{\max} \end{array} \right\}. \tag{26}$$

Combining the two regions $\mathcal{M}_{y^{\pm}}^{(w)}$ results in the unified region $\mathcal{M}^{(w)} = \mathcal{M}_{y^{-}}^{(w)} \cup \mathcal{M}_{y^{+}}^{(w)}$ for each side-wall.

2) Floor and Ceiling: Varying the distance paramater d_z given in (11) from $d_w^{\min}$ to $d_w^{\max}$ translates into a curve segment in polar-coordinate space described by

$$\underbrace{\sqrt{\rho_o^2 + (2d_w^{\min})^2}}_{\rho_w^{\min}} \leq \rho \leq \underbrace{\sqrt{\rho_o^2 + (2d_w^{\max})^2}}_{\rho_w^{\max}} \tag{27}$$

and

$$\theta(\rho) = \arccos\left(\frac{\rho_o \cos\theta_o}{\rho}\right), \tag{28}$$

yielding the bounded regions $\mathcal{M}^{(w)}$ $(w = 5, 6)$ on the DOA map containing the first-order floor and ceiling reflections:

$$\mathcal{M}^{(w)} = \left\{ (\rho, \theta) \ : \begin{array}{l} \rho_w^{\min} \leq \rho \leq \rho_w^{\max} \\ \theta(\rho) - \xi_\theta \leq \theta \leq \theta(\rho) + \xi_\theta \end{array} \right\}, \tag{29}$$

where a small tolerance ξ_θ is introduced on the angle $\theta(\rho)$ to convert each $\mathcal{M}^{(w)}$ from a curve segment to a narrow-width region.

C. Candidates for First-Order Reflections

The set of candidate peaks within the bounded region $\mathcal{M}^{(w)}$ corresponding to the wth wall is determined by intersecting $\mathcal{M}^{(w)}$ with the set $\overline{\mathcal{Q}}_{\mathcal{K}}$, which only includes the $\mathcal{K}$ largest peaks from $\overline{\mathcal{Q}}$:

$$\overline{\mathcal{Q}}^{(w)} = \mathcal{M}^{(w)} \cap \overline{\mathcal{Q}}_{\mathcal{K}}. \tag{30}$$

If still $\overline{\mathcal{Q}}^{(w)} = \emptyset$, then the search for a peak among the remaining elements of $\overline{\mathcal{Q}}$ is continued until $\overline{\mathcal{Q}}^{(w)}$ has at least one element. If the cardinality of $\overline{\mathcal{Q}}^{(w)}$ is greater than a pre-specified number $\widetilde{\mathcal{K}}$, then only the first $\widetilde{\mathcal{K}}$ elements are kept in $\overline{\mathcal{Q}}^{(w)}$ and the rest are discarded to avoid an exhaustive search.

To account for the geometrical ambiguity introduced by the linear array, $\overline{\mathcal{Q}}^{(w)}$ is subdivided back into two subsets as $\overline{\mathcal{Q}}_{y^{\pm}}^{(w)} = \overline{\mathcal{Q}}^{(w)} \cap \mathcal{M}_{y^{\pm}}^{(w)}$ for the side-walls $(w = 1, \ldots, 4)$. Given the real-microphone position $\mathbf{r}_o$, each pair (ρ, θ) in $\overline{\mathcal{Q}}_{y^{\pm}}^{(w)}$ are translated into their corresponding wall parameters $(\mathbf{v}_{y^{\pm}}^{(w)}, d_{y^{\pm}}^{(w)})$ using (3), (6) and (7) and stored in a unified wall-parameter set denoted by $\mathcal{P}^{(w)}$. For floor and ceiling $(w = 5, 6)$, the wall-parameter sets $\mathcal{P}^{(w)}$ are obtained directly from the peak set $\overline{\mathcal{Q}}^{(w)}$.

VI. Room Geometry Inference

A. Room Geometry Candidates

Considering all possible combinations of individual wall parameters, the room geometry candidates are obtained via the Cartesian product as

$$\begin{aligned} \mathcal{G} &= \mathcal{P}^{(1)} \times \mathcal{P}^{(2)} \times \cdots \times \mathcal{P}^{(6)} \\ &= \{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_J\}, \end{aligned} \tag{31}$$

where $J = |\mathcal{G}|$ is the total number of candidate room geometries and the jth candidate described by the parameters $(\mathbf{v}_{w,j}, d_{w,j})$ for each wall is given by

$$\mathcal{G}_j = \{(\mathbf{v}_{1,j}, d_{1,j})\,; \ldots\,; (\mathbf{v}_{6,j}, d_{6,j})\}. \tag{32}$$

Considering the $z = 0$ plane, on which the real- and image-microphones lie, side-walls can be represented by the lines of equation $a_{w,j}x + b_{w,j}y + c_{w,j} = 0$, corresponding to the homogeneous vector [35]

$$\mathbf{l}_{w,j} = \begin{bmatrix} a_{w,j} \\ b_{w,j} \\ c_{w,j} \end{bmatrix} = \begin{bmatrix} [\mathbf{v}_{w,j}]_1 \\ [\mathbf{v}_{w,j}]_2 \\ d_{w,j} \end{bmatrix}, \tag{33}$$

where $[\mathbf{v}]_k$ denotes the kth component of the vector $\mathbf{v}$. The 2D Cartesian coordinates $\mathbf{p}_{1,j}, \ldots \mathbf{p}_{4,j}$ of the four corners are computed by intersecting the pairs of lines[1] $(\mathbf{l}_{1,j}, \mathbf{l}_{2,j})$, $(\mathbf{l}_{2,j}, \mathbf{l}_{3,j})$, $(\mathbf{l}_{3,j}, \mathbf{l}_{4,j})$ and $(\mathbf{l}_{4,j}, \mathbf{l}_{1,j})$. The floor map is then described by the set of line segments

$$\mathcal{F}_j = \{\overline{\mathbf{p}_{1,j}\mathbf{p}_{2,j}}\,, \overline{\mathbf{p}_{2,j}\mathbf{p}_{3,j}}\,, \overline{\mathbf{p}_{3,j}\mathbf{p}_{4,j}}\,, \overline{\mathbf{p}_{4,j}\mathbf{p}_{1,j}}\}. \tag{34}$$

The candidate room height is simply given by

$$H_j = d_{5,j} + d_{6,j}. \tag{35}$$

Arising from the up-down ambiguity, the projection of the set of higher-order image-microphone positions of two candidates j' and j'' onto the DOA map would be the same if $d_{5,j'} = d_{6,j''}$ and $d_{6,j'} = d_{5,j''}$, which may result in an "upside-down" inferred room geometry. Therefore, the linear loudspeaker array and single microphone are always assumed to be positioned closer to the floor, meaning that any given candidate $\mathcal{G}_j$ with $d_{5,j} > d_{6,j}$ is discarded from the candidate set $\mathcal{G}$ to tackle the up-down ambiguity.

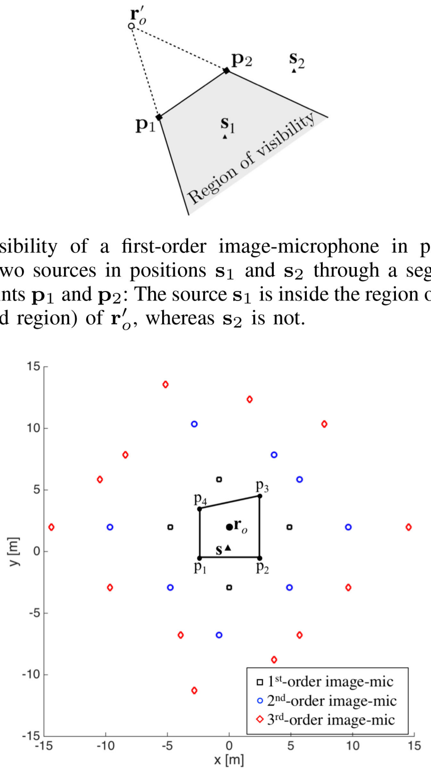
To prevent unrealistically low estimates for the room height particularly in real-world scenarios with noisy measurements, a minimum room height level is also introduced as an additional constraint such that when $H_j < H_{\min}$ for a candidate $\mathcal{G}_j$, it is also omitted from $\mathcal{G}$. If no candidate is left in $\mathcal{G}$ satisfying the minimum height constraint, then the search among the remaining elements of $\overline{\mathcal{Q}}$ is continued until at least a peak-pair for floor and ceiling satisfies $H_{\min}$.

B. Beam Tracing and Cost Function

In the final step, a candidate from the set $\mathcal{G}$ is selected as the inferred room geometry, comparing the estimated higher-order reflections with the actual peaks on the DOA map remaining after the exclusion of the peak-set corresponding to first-order reflections. In more detail, for each room geometry candidate, the higher-order image-microphone positions are first estimated via the beam-tracing method described in [27]–[29] from the

[1] Given a pair of lines $\mathbf{l}_1 = [a_1, b_1, c_1]^T$ and $\mathbf{l}_2 = [a_2, b_2, c_2]^T$, their intersection is computed in homogeneous coordinates as the cross product $\mathbf{l}_1 \times \mathbf{l}_2$ [35]. It is easy to verify that the corresponding Cartesian coordinates are given by

$$\mathbf{p} = (x, y) = \left(\frac{b_1c_2 - b_2c_1}{a_1b_2 - a_2b_1}, \frac{a_2c_1 - a_1c_2}{a_1b_2 - a_2b_1}\right).$$



(a) Visibility of a first-order image-microphone in position $\mathbf{r}'_0$ from two sources in positions $\mathbf{s}_1$ and $\mathbf{s}_2$ through a segment with endpoints $\mathbf{p}_1$ and $\mathbf{p}_2$: The source $\mathbf{s}_1$ is inside the region of visibility (shaded region) of $\mathbf{r}'_o$, whereas $\mathbf{s}_2$ is not.

(b) An illustration of the set of visible image-microphone positions obtained through the beam-tracing method for a given room.

Fig. 6. Visibility of an image-microphone and an example of a set of visible image-microphones obtained via beam tracing.

candidate first-order reflections, and then projected onto the 2D polar-coordinate space as illustrated in the rightmost column of Fig. 5. Subsequently, a cost function is computed by summing the amplitudes of grid points on the DOA map corresponding to the estimated higher-order image-microphone positions. Finally, the candidate that achieves the highest score among all the others is determined as the inferred room geometry.

1) Beam Tracing: For a source-receiver pair along with a set of line segments describing a 2D acoustic environment, beam tracing [28] enables the computation of all image-microphone positions that are visible from the source using the image method [36] up to a predetermined order of reflections, in which the visibility from the sources is evaluated by tracing the beam splitting/branching process resulting from multiple reflections on the line segments. The concept of visibility is exemplified in Fig. 6(a), where the region of visibility of an image-microphone at $\mathbf{r}'_o$ through the segment $\overline{\mathbf{p}_1\mathbf{p}_2}$ corresponds to a beam (gray-shaded region), and the image microphone is visible from the source at $\mathbf{s}_1$ but not from $\mathbf{s}_2$. As discussed in [28], [29], the beam-tracing method originally developed for planar (2D) geometries can be easily extended to 2D × 1D geometry. This is accomplished by mirroring the set of the resulting coplanar image-microphones with respect to the planes corresponding to floor and ceiling, and repeating this procedure until reaching the maximum reflection order.

To estimate the higher-order propagation paths for the room-geometry candidate j, the beam-tracing process may be defined in the form of a function as

$$\mathcal{I}_j = B\left(\mathcal{F}_j, H_j, \mathbf{s}, \mathbf{r}_o, \kappa\right), \tag{36}$$

which yields the output set $\mathcal{I}_j$ containing the 3D Cartesian coordinates of all image-microphones visible from the array center $\mathbf{s} = [0, 0, 0]^T$ given the real microphone position $\mathbf{r}_o$ and the maximum order of reflections of interest, denoted by κ. An illustration of the beam-tracing output up to the 3rd-order image-microphones is depicted in Fig. 6(b) for a room with a trapezoidal floor map.

2) Cost Function: To evaluate the agreement between the higher-order reflections estimated via beam tracing and the actual peaks on the DOA map for the room-geometry candidate j, a cost function is defined as

$$\Psi(\mathring{\mathcal{I}}_j) = \sum_{(\rho_j,\theta_j)\in\mathring{\mathcal{I}}_j} \max_{\substack{|\rho-\rho_j|<\tau_\rho \\ |\theta-\theta_j|<\tau_\theta}} X_{\rho,\theta}, \quad j \in \{1, 2, \ldots, J\}, \tag{37}$$

where $\mathring{\mathcal{I}}_j$ is the set of image-microphone positions given in the polar-coordinate space, obtained by applying (3) to the elements of $\mathcal{I}_j$, and the thresholds τ_ρ and τ_θ describe a rectangular region on the DOA map centered at each position $(\rho_j, \theta_j) \in \mathcal{I}_j$, within which the local maximum is then determined.

The candidate that attains the highest score is selected as the final estimate:

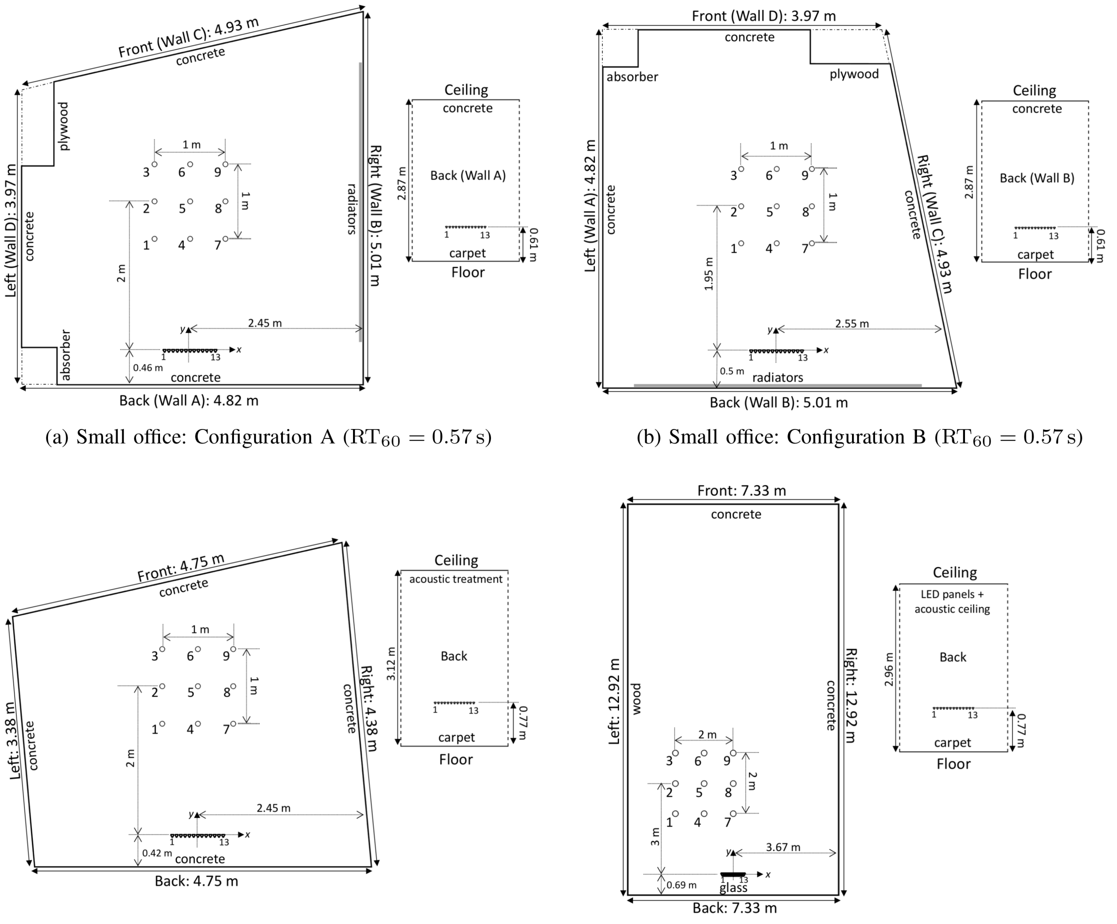
$$\hat{j} = \arg\max_j \Psi(\mathring{\mathcal{I}}_j), \tag{38}$$

and the room geometry is then inferred from the set of wall parameters $\mathcal{G}_{\hat{j}}$.

VII. Performance Evaluation

A. Experiments

The proposed method was tested in three real rooms (two configurations in a small office, a laboratory room and a lecture hall) as illustrated with top- and side-views in Fig. 7. The measurement setup consisted of a 13-element uniform linear array of 2-inch full-range drivers spaced apart by 6 cm and an omni-directional microphone of type *Microtech Gefell M373*. The loudspeaker array was positioned close and parallel to the back-wall as a typical office/home layout and the microphone was placed in front of it at a total of nine positions according to an equally-spaced 3×3 grid. The individual transmission paths between each loudspeaker in the array and the microphone were measured using a logarithmic sweep with a time-length of 5 s starting at 150 Hz and reaching its end at 20 kHz, and RIRs were obtained using a standard deconvolution scheme. A loopback measurement was used to free the measurements from the hardware latency of the system such that any initial delay in RIRs was associated with the acoustic propagation of the direct path. Additionally, the reverberation times were measured in all three rooms according to the ISO standard [37]. The sampling frequency was 48 kHz, and the air temperature was set to 20 °C corresponding to a speed of sound of approximately 343 m/s.



(a) Small office: Configuration A ($\mathrm{RT}_{60} = 0.57\,\mathrm{s}$)

(b) Small office: Configuration B ($\mathrm{RT}_{60} = 0.57\,\mathrm{s}$)

(c) Laboratory room ($\mathrm{RT}_{60} = 0.7\,\mathrm{s}$)

(d) Lecture hall ($\mathrm{RT}_{60} = 0.92\,\mathrm{s}$)

Fig. 7. The experimental setups used for the evaluation of the proposed 3D RGI method: RIRs were measured at nine different microphone positions in a small room with two configurations, in a laboratory room and a lecture hall with the measured reverberation times reported in parentheses.

1) Small Office: As depicted in Figs. 7(a) and 7(b), the room had four flat concrete side-walls featuring a non-rectangular floor map, a floor covered with a thin carpet and a concrete ceiling parallel to the floor, satisfying the 2D × 1D geometry assumption. A radiator and a cable shaft were located on one side of the room (indicated by the label "radiators"), both causing scattering of incident sound, and a plywood closet was located at one of the corners. The experiments were performed in two different configurations named as "Configuration A" and "Configuration B". Identical walls were labelled with the same name (e.g., *Wall A*, *Wall B*, etc.) in both cases, while their roles (e.g., *Front*, *Left*, etc.) changed depending on the configuration. The reverberation time was measured to be $\mathrm{RT}_{60} = 0.57\,\mathrm{s}$.

2) Laboratory Room: As illustrated in Fig. 7(c), the room had a trapezoidal floor map with all four walls made of concrete and a floor covered with a thin layer of carpet. The ceiling was, however, acoustically treated with a layer of porous absorption behind a perforated ceiling panel, attenuating acoustic reflections. Furthermore, just below the ceiling, an air ventilation duct protruding into the room additionally had an impact on reflections off of the ceiling. The reverberation time was measured to be $\mathrm{RT}_{60} = 0.7\,\mathrm{s}$.

3) Lecture Hall: As shown in Fig. 7(d), the lecture hall was the largest of the three rooms, while it was also the only one with a rectangular floor-map. The back-wall was actually a full glass window with steel frames, creating a non-uniform reflective surface, the left-wall was a wooden room-partitioning panel, and front- and right- walls were made of concrete. The floor was covered with a thin layer of carpet, whereas the ceiling featured another non-uniform surface consisting of a mixture of LED and acoustic-ceiling panels. The reverberation time was measured to be $\mathrm{RT}_{60} = 0.92\,\mathrm{s}$.

B. Simulations

A replica of the two configurations in the small office excluding the absorber and the plywood cabinet was simulated using the same loudspeaker and microphone setup (but assuming omnidirectional responses) to compare the performance of the RGI algorithm between the experimental and simulated data. The simulated RIRs were generated via the beam-tracing method up to the tenth-order of reflections [28], [29]. To test the proposed method at varying reverberation times, RIRs were simulated at RT_{60}s of 0.2 s, 0.6 s and 1.0 s using Configuration A, where the wall attenuation coefficients were computed using the Sabine's formula [24]. In addition, RIRs were generated using both configurations to evaluate the robustness of the RGI algorithm to varying wall characteristics by setting the attenuation coefficient of the wall along which the radiators were installed (labelled as "Wall B," playing the role of a "right-wall" in Configuration A, and a "back-wall" in Configuration B) to $\alpha_w = 0.1$ and $\alpha_w = 0.5$, while keeping the attenuation coefficients of the remaining walls fixed at $\alpha_w = 0.9$.

C. Parameter Selection

1) DOA Mapping: Each RIR measurement was truncated at the sample corresponding to the propagation distance of $\rho_{\max} = 15\,\text{m}$ for the small office and laboratory room, resulting in $N = \lfloor \rho_{\max} \cdot f_s/c \rceil = 2099$, and $\rho_{\max} = 50\,\text{m}$ for the lecture hall, yielding $N = 6997$. These values were chosen large enough for each room to accommodate higher-order reflections on the DOA map for beam tracing. The DOA-map resolution along the ρ-axis was set to be the same as the measured RIRs i.e., $M = N$. The angular grid was obtained by uniformly sampling over $[0°, 180°]$ with a resolution of $1°$, leading to a grid length of $I = 181$.

The smoothing filter $\mathbf{g}^{(l)}_{\rho_m,\theta_i}$ used for the generation of the matrix $\mathbf{A}$ was sampled from a rectangular-windowed Laplace probability density function as an ad-hoc choice to replace the 1D Gaussian filter with the aim of peak-sharpening and hence a reduction in blurring when compared to a Gaussian filter:

$$g^{(l)}_{\rho_m,\theta_i}(n) = \begin{cases} \frac{1}{2\sigma} e^{-\left|n - f_s \cdot d^{(l)}_{\rho_m,\theta_i}/c\right|/\sigma} & \text{if } \mathcal{N}_l \leq n \leq \mathcal{N}_u \\ 0 & \text{otherwise} \end{cases} \tag{39}$$

where σ is the scale parameter and the rectangular window is described by the lower bound $\mathcal{N}_l = \max(0, -N_{win} + \lfloor f_s \cdot d^{(l)}_{\rho_m,\theta_i}/c + 1)$ and the upper bound $\mathcal{N}_u = \min(N-1, N_{win} + \lfloor f_s \cdot d^{(l)}_{\rho_m,\theta_i}/c \rfloor)$, where $N_{win} \in \mathbb{Z}^+$ and $\lfloor \cdot \rfloor$ refers to the flooring to the closest integer value operation. The choice for the filter parameters were $\sigma = 1$ and $N_{win} = 4$, resulting in a narrow-width filter with a sharp peak. This procedure may also be viewed as the selection of a filter that could generate a point spread function for the imaging system, i.e., the local impulse response to a point source [38] with desired properties. Alternatively, fractional-delay filters as described in [32] could be used for the same purpose, but this would result in the matrix $\mathbf{A}$ to have negative-valued elements, which could complicate the map estimation with a non-negativity constraint as the measured RIRs also included the loudspeaker driver responses, causing the peaks in RIRs to follow a pattern quite different to that of a sinc function [39].

The Elastic Net regularization parameters, (λ, β), were empirically tuned to achieve an appropriate trade-off between the model agreement and locally-smooth sparseness: $\lambda = 0.1$ and $\beta = 0.05$ for the small office and laboratory room, and $\lambda = 0.01$ and $\beta = 0.05$ for the lecture hall, where a smaller λ was used for the lecture hall to also recover the peaks appearing much later in the RIRs (as the lecture hall being a much larger room than the other two) because LASSO tended to reduce the amplitude of these peaks more than the early ones at a given λ due to the distance compensation arising from the spherical-wave propagation model. The relative error threshold used as the stopping criterion for the coordinate descent algorithm was set to be $\tau = 10^{-3}$, resulting in around 30 iterations. MATLAB was run on a computer with a 3.5 GHz Intel Core i7 processor and a 16 GB RAM, solving the Elastic Net in ~15 minutes for the small office and the laboratory room, and in ~30 minutes for the lecture hall.

TABLE I
PARAMETERS USED FOR THE WALL CONSTRAINTS

	Small office / Lab. room			Lecture hall		
Wall	$d_w^{\min}$	$d_w^{\max}$	$\Delta\varphi_w$	$d_w^{\min}$	$d_w^{\max}$	$\Delta\varphi_w$
Back	0.2 m	1.0 m	15°	0.2 m	1.0 m	15°
Right	1.5 m	3.0 m	15°	1.5 m	6.0 m	15°
Front	3.0 m	6.0 m	15°	9.0 m	14.0 m	15°
Left	1.5 m	3.0 m	15°	1.5 m	6.0 m	15°
Floor	0.5 m	1.5 m	-	0.5 m	1.5 m	-
Ceiling	0.7 m	2.7 m	-	0.7 m	2.7 m	-

For the peak detection and pruning process, the values defining the size of the neighborhood were empirically determined as $T_\theta = 5°$ and $T_\rho = 0.034\,\text{m}$ (equivalently a time difference of 0.1 ms) to make a reasonable trade-off between reducing the possibility of missing any closely positioned peaks and incorrectly detecting some image artifacts appearing with a small amplitude on the DOA map as potential peaks corresponding to acoustic reflections.

2) Map Segmentation: The wall constraints used for the generation of bounded regions, $d_m^{\min}$, $d_m^{\max}$ and $\Delta\varphi_m$ were imposed as two separate sets for the small rooms (small office and laboratory room) and the lecture hall as listed in Table I. The geometrical bounds selected for the small rooms were intended to be general enough for a wide range of small offices and living rooms, which may also be treated as very rough estimates that could potentially be given by a consumer through visual inspection of the actual scene in a commercial scenario. These bounds, however, had to be relaxed and adapted for the lecture hall, as the RGI algorithm was not initially designed to work in such large rooms.

The search inside each bounded region for the peak candidates for first-order reflections was initialized with the $\mathcal{K} = 50$ largest peaks from $\overline{\mathcal{Q}}$, and a maximum of $\widetilde{\mathcal{K}} = 10$ peaks were allowed to be candidates within each bounded region.

3) Room Geometry Inference: The minimum room height constraint was set to $H_{\min} = 2.2\,\text{m}$, which should be a reasonable value considering a typical human height range, and also again may be given as a rough user input in a commercial scenario as described above.

The beam tracing was run with the maximum reflection order to be $\kappa = 3$. The thresholds used for the evaluation of the cost function were as $\tau_\rho = 0.02\,\text{m}$ and $\tau_\theta = 2°$. These values were found experimentally to be large enough to guarantee a sufficient level of robustness against errors in the beam-tracing output (mainly caused by uncertainties in the positions of the array elements and microphone, as well as temporal- and

TABLE II
EXPERIMENTS: INDIVIDUAL WALL ESTIMATION ERROR

Error	Room	Back-wall	Right-wall	Front-wall	Left-wall	Floor	Ceiling
$\epsilon_{w,d}$ [cm]	Configuration A	8.805 ± 9.089	6.501 ± 9.878	0.518 ± 0.684	2.662 ± 2.293	0.673 ± 0.317	1.494 ± 0.255
	Configuration B	9.087 ± 6.829	2.072 ± 0.744	37.567 ± 50.427	5.623 ± 5.843	0.902 ± 0.320	1.777 ± 0.130
	Laboratory room	25.045 ± 15.454	3.779 ± 3.344	8.584 ± 2.042	8.293 ± 2.303	1.426 ± 3.704	4.018 ± 7.770
	Lecture hall	35.120 ± 6.453	7.242 ± 5.375	21.376 ± 11.353	3.703 ± 3.094	0.537 ± 0.406	2.913 ± 0.797
$\epsilon_{w,\theta}$ [°]	Configuration A	0.602 ± 0.673	2.792 ± 2.709	0.574 ± 0.333	1.065 ± 0.687	-	-
	Configuration B	1.112 ± 0.976	0.393 ± 0.289	1.010 ± 0.717	2.612 ± 1.892	-	-
	Laboratory room	1.263 ± 1.362	1.346 ± 0.789	0.757 ± 0.617	1.828 ± 1.793	-	-
	Lecture hall	0.472 ± 0.395	2.356 ± 1.547	0.765 ± 0.582	1.152 ± 0.692	-	-

Mean±std (standard deviation) computed over nine measurement positions.

TABLE III
EXPERIMENTS: RGI ERROR

Room	E_d [cm]	E_θ [°]
Configuration A	5.809 ± 4.171	1.681 ± 1.272
Configuration B	17.595 ± 19.409	1.666 ± 0.870
Laboratory room	12.521 ± 5.443	1.628 ± 0.773
Lecture hall	17.720 ± 3.565	1.508 ± 0.668

Mean±std (standard deviation) computed over nine measurement positions.

spatial-sampling artifacts). At the same time, they were sufficiently small to avoid the presence of multiple local maxima falling within the neighborhood of the same candidate image-microphone position.

D. Evaluation Metrics

To evaluate the estimation accuracy for individual walls, the deviation of the estimated pair of wall parameters $(\hat{\mathbf{v}}_w, \hat{d}_w)$ from the ground-truth values $(\mathbf{v}_w, d_w)$ is computed in terms of the distance error $\epsilon_{w,d}$ and orientation error $\epsilon_{w,\theta}$ given by

$$\epsilon_{w,d} = |d_w - \hat{d}_w| \quad \text{and} \quad \epsilon_{w,\theta} = \arccos\langle \mathbf{v}_w, \hat{\mathbf{v}}_w \rangle, \tag{40}$$

respectively. Please note that $\epsilon_{w,\theta} = 0$ for the floor and ceiling under the separable room geometry assumption, and hence is considered only for the side-wall estimates.

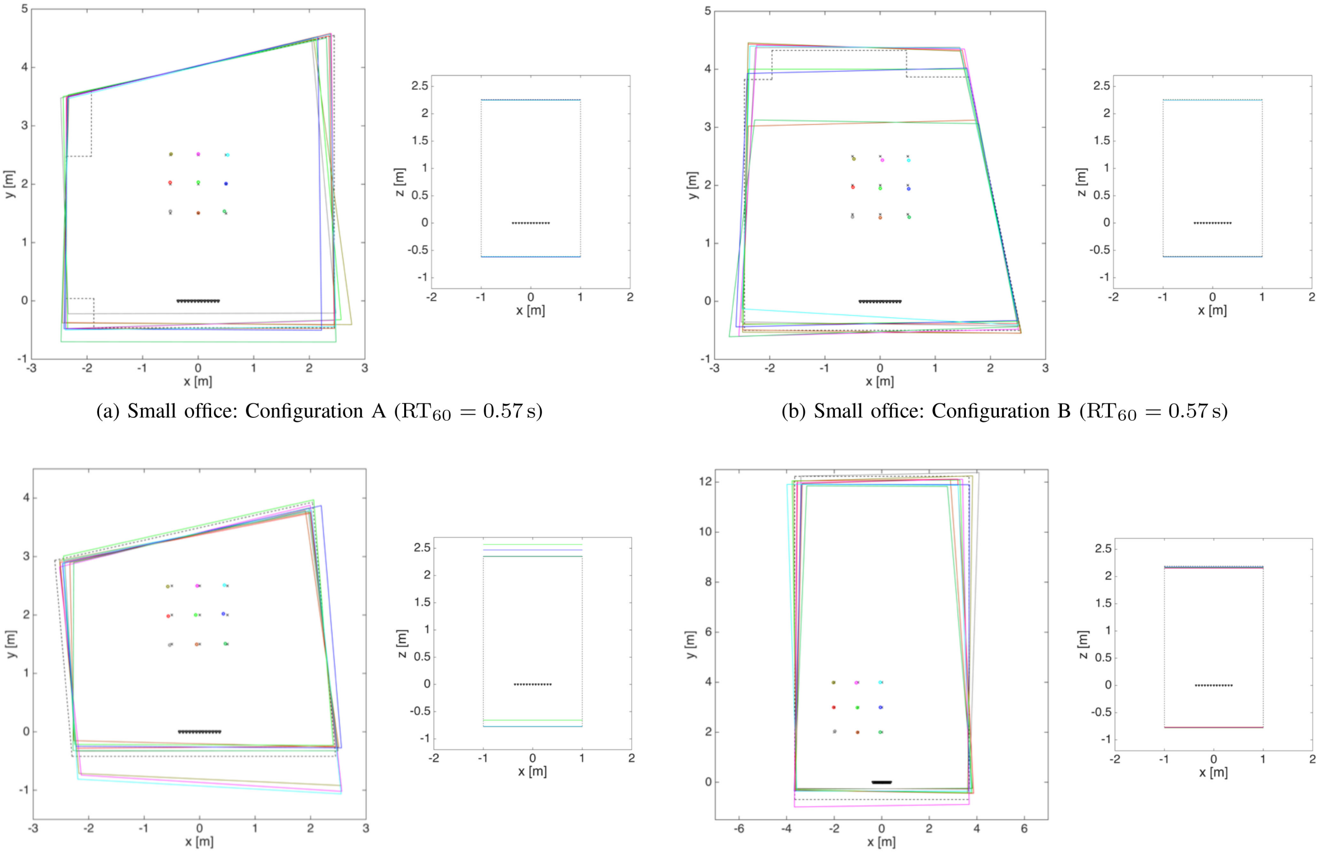
To jointly assess the accuracy of the inferred room geometry based on all estimated walls, the root mean square distance error E_d and the root mean square orientation error E_θ are defined as

$$E_d = \sqrt{\frac{1}{6} \sum\nolimits_{w=1}^{6} \epsilon_{w,d}^2}, \quad E_\theta = \sqrt{\frac{1}{4} \sum\nolimits_{w=1}^{4} \epsilon_{w,\theta}^2}, \tag{41}$$

where E_θ is computed only over the side-wall estimates, since $\epsilon_{w,\theta} = 0$ by assumption as mentioned above.

E. Results

The performance evaluation results for the experimental data are presented in Table II using the metrics $\epsilon_{w,d}$ and $\epsilon_{w,\theta}$, and in Table III using the metrics E_d and E_θ. In both tables, the mean and standard deviation of each metric computed over nine measurements positions are reported. In addition, the inferred floor maps and room height at nine positions are plotted in Fig. 8 for all four setups in comparison to the actual layout along with the estimated microphone positions. The mean-error values for individual wall estimates $\mu(\epsilon_{w,d}) < 38\,\text{cm}$ and $\mu(\epsilon_{w,\theta}) < 3^\circ$ indicate that the proposed method has achieved a reasonable accuracy across the measurement positions. Despite the geometrical ambiguity arising from using a 1D array, the RGI mean-error values $\mu(E_d) < 18\,\text{cm}$ and $\mu(E_\theta) < 2^\circ$ are numerically comparable to those achieved by the techniques in the literature operating both with 2D and 3D array configurations in rooms with similar dimensions and acoustic characteristics [10], [12], [13], [16], [22]. The fact that the wall with a long radiator in the small office was not a proper reflector can be observed in the error results and estimated floor maps, particularly in Configuration A, where it was in the role of a right-wall. The effect of the speaker directivity on the individual wall estimates can clearly be seen from the distance-error comparison between the back-wall and others, where in all four setups, the back-wall was behind the loudspeaker array. The most severe case with a very high error occurred in the laboratory room at two positions, where the peaks corresponding to the back-wall were not even picked during the peak selection for the first-order reflections due to having very weak amplitudes. The exceptional case resulted with the highest error took place during the estimation of the front-wall in Configuration B, where at two positions, the estimated front-wall was aligned with the absorber and plywood closet, and in two other positions, the peak corresponding to the ceiling estimate was also selected as the front-wall estimate, as the RGI algorithm was allowed to select the same peak for multiple walls. If the loudspeaker array is positioned parallel to a side-wall, then the associated peaks line up along the same path on the DOA map as the peaks corresponding to higher-order floor and/or ceiling reflections as a result of the lack of spatial diversity in the linear array. Therefore, these peaks were inaccurately associated with the front-wall by the RGI algorithm during the beam tracing and cost function step due to the actual higher-reflections from the front-wall potentially shifted by the plywood cabinet (a "non-flat" wall scenario) and attenuated by the absorber. It was found out that if each peak candidate was restricted to a single-wall, the algorithm was able to select the correct peak for the front-wall at these two positions. For the same reason, it was also observed that if the minimum height constraint $H_{\min}$ was not introduced, the RGI algorithm would select a peak appearing earlier than the actual ceiling reflection on the DOA map as the ceiling candidate, resulting in an unrealistically low height estimate, particularly at two positions in Configuration B and



(a) Small office: Configuration A ($\mathrm{RT}_{60} = 0.57\,\mathrm{s}$)

(b) Small office: Configuration B ($\mathrm{RT}_{60} = 0.57\,\mathrm{s}$)

(c) Laboratory room ($\mathrm{RT}_{60} = 0.7\,\mathrm{s}$)

(d) Lecture hall ($\mathrm{RT}_{60} = 0.92\,\mathrm{s}$)

Fig. 8. Room geometry inference results: The floor-map and floor/ceiling height inferred at all nine positions in four setups along with the estimated microphone positions.

at the majority of grid positions in the lecture hall because of the acoustic ceiling attenuating the reflections. However, even with $H_{\min}$, the room height estimate was still incorrect at two positions in the laboratory room due to the duct and acoustic treatment across the ceiling. Besides, the measurement grid in the lecture hall was intentionally positioned closer to the left-wall, resulting in the right-wall estimates to be slightly worse as expected. Since the front-wall was located far away from the measurement positions, it was estimated with a relatively high error. It was also found out that if a wide range was used for the front-wall distance constraint, the RGI algorithm tended to select one of the very earlier peaks appearing along the same path on the DOA map as the front-wall peak resulting from the geometrical ambiguity similar to the case described above for Configuration B.

TABLE IV
SIMULATIONS: RGI ERROR

Room: Small office	E_d [cm]	E_θ [°]
Configuration A: $\mathrm{RT}_{60} = 0.2$ s	0.706 ± 0.372	0.499 ± 0.295
Configuration A: $\mathrm{RT}_{60} = 0.6$ s	3.924 ± 6.701	1.467 ± 1.959
Configuration A: $\mathrm{RT}_{60} = 1.0$ s	2.411 ± 3.810	1.070 ± 1.458
Configuration A: $\alpha_{\mathrm{Right}} = 0.1$	5.116 ± 6.415	1.686 ± 2.202
Configuration A: $\alpha_{\mathrm{Right}} = 0.5$	1.531 ± 2.390	0.753 ± 0.675
Configuration B: $\alpha_{\mathrm{Back}} = 0.1$	4.309 ± 5.377	0.567 ± 0.297
Configuration B: $\alpha_{\mathrm{Back}} = 0.5$	0.720 ± 0.236	0.572 ± 0.293

Mean±std (standard deviation) computed over nine measurement positions.

The RGI error results for the simulated data are presented in Table IV. The RGI algorithm showed better performance with the simulated data than in experiments. This was not surprising, since the simulations represented an idealized empty-room scenario, where the walls were perfectly flat with no irregularities, loudspeakers and microphone were perfectly omnidirectional and there were no additional acoustic phenomena such as scattering or diffraction. There was a slight increase in RGI error with longer reverberation times because the walls became more reflective with increasing RT_{60}, which in turn yielded larger peaks appearing in the later part of the DOA map, causing the algorithm to miss the correct peaks in some cases during the peak selection step for first-order reflection candidates. This was due to these peaks not being included in the set of $\mathcal{K} = 50$ largest peaks, where $\mathcal{K}$ was actually empirically set based on the experimental data. As expected, the performance degraded with a less reflective wall, where the RGI algorithm usually selected a random peak from the corresponding bounded region on the DOA map, as the actual peak associated with the wall was very weak. The distance-error for Wall B in the simulated replica of the small office was $\epsilon_{w,d}$ [cm] : 12.112 ± 16.020 when $\alpha_{\mathrm{Right}} = 0.1$ and $\epsilon_{w,d}$ [cm] :

3.207 ± 6.048 when $\alpha_{\text{Right}} = 0.5$ in Configuration A (role: right-wall), and $\epsilon_{w,d}$ [cm] : 9.635 ± 13.812 when $\alpha_{\text{Back}} = 0.1$ and $\epsilon_{w,d}$ [cm] : 0.456 ± 0.145 when $\alpha_{\text{Back}} = 0.5$ in Configuration B (role: back-wall).

F. Discussion and Future Work

The adoption of spherical wave propagation in the physical model enables a more accurate description of the acoustic wavefronts in the near-field. Assuming a plane-wave propagation model would particularly be problematic when the microphone was positioned close to the loudspeaker array with respect to the array size because this would cause the peaks corresponding to the direct path and first-order reflections to appear as smeared and angularly shifted on the DOA map, increasing the wall estimation errors noticeably. A similar problem is also reported in [22], where the linear Radon transform is applied on RIRs for TOA disambiguation. As one would expect, the use of a spherical wave propagation model may produce artificial peaks on the DOA map at the high values of ρ due to the noise in RIRs boosted by the distance compensation. However, this is mitigated through the quadratic penalty in the optimization cost function reducing the noise with a proper choice of the regularization parameters. At the moment, these parameters are manually tuned, since it can be quite challenging to achieve an appropriate trade-off between the model agreement and the locally-smooth sparseness. Nevertheless, the algorithm would highly benefit from the development of an automatic parameter tuning procedure in the future. Another potential disadvantage of the optimization framework used in map generation may be the computational complexity. However, the sparse matrix structure and the lack of need for any inverse operation during the implementation of the coordinate descent algorithm makes the optimization step quite feasible in terms of computational time. The peak detection and pruning technique used in this work is rather a simple approach, meaning that it is incapable of differentiating the peaks that are weak in amplitude but actually represent an acoustic reflection from image artifacts caused by model discrepancies and/or noise. An example of such limitation can be seen in the rightmost column of Fig. 4 especially around the direct-path peak. The use of a more advanced scheme such as the morphology-based peak-detection scheme used in [20] may bring some significant improvement, simplifying the selection of peak candidates for the first-order reflections on the segmented map, essentially when a reflection from a wall appears as rather a weak peak on the DOA map due to absorption and/or transducer directivity.

Map segmentation involves the most challenging part of the proposed RGI method, since during this step, the geometrical ambiguity is tackled by an empirical approach built upon the generation of bounded regions on the DOA map with respect to the pre-defined constraints on the wall dimensions and orientations. There is also room for improvement regarding the selection of peak candidates for the first-order wall reflections, as the current algorithm may fail to include the peaks associated with weakly reflective surfaces. Thus, another potential advancement should include the replacement of the semisupervised approach for the generation of wall-candidate sets with a fully automated scheme. The current RGI algorithm is expected to work in rooms with more than four side-walls and possibly with more inclination, provided that the constraints are given for all the walls. A tilted ceiling scenario may also be handled by the algorithm, but this would require a modification in the beam-tracing step. However, similar to many existing RGI algorithms, all these can become possible only if all the first-order reflections exist on the DOA map. Therefore, the current algorithm will either fail or select a random peak from the bounded regions when a first-order reflection is absent from the DOA map due to a very absorbent boundary, invisibility or occlusion.

While the cost function used for RGI yields reliable results in relatively empty rooms, its dependence on the position and amplitude of higher-order reflections on the DOA map may also easily lead to inaccurate estimates, particularly for rooms with many furniture items. Diffuse reflections may also be another factor degrading the estimation accuracy, as they result in sets of smeared peaks on the DOA map, generating higher-order reflections that are also scattered and much more weakened in magnitude. To tackle these issues, the DOA map itself may actually be very useful, since it may also provide more acoustical information about the walls, other reflective surfaces and potential furniture in the room. The generation of DOA maps on bandpass-filtered RIRs at multiple frequencies may enable the frequency-dependent analysis of individual peaks, which may then be used for a more robust and comprehensive characterization of acoustic reflectors within a room.

The proposed RGI method operates with a synchronized setup to be able to directly estimate DOAs from multiple RIRs without the need to detect TOA in each microphone-speaker pair individually. When synchronization is not possible in advance, auto-calibration techniques such as [16], [17] may be adopted as an initial step for measurement-position estimation. The developed approach requires the measurement microphone to be placed in front of the array to avoid geometrical ambiguity introduced by the linear speaker array. However, this may be regarded as a plausible condition in a typical home/office setting.

VIII. Conclusion

We have proposed a 3D RGI methodology for the identification of wall reflectors in convex-shaped rooms from acoustic measurements collected with a linear loudspeaker array and a single microphone. RIRs are translated into a high resolution 2D DOA map, on which salient peaks represent distinct acoustic reflections as a function of the propagation distance and DOA angle with respect to the speaker array. Assuming a 2D × 1D geometry, relaxed geometrical wall bounds are used for the segmentation of the DOA map into six regions to localize the first-order reflections corresponding to four side-walls, floor and ceiling. The 3D wall parameters are determined as those maximizing a cost function evaluated on the DOA map using the higher-order image-microphone positions estimated via beam tracing from the first-order reflections, among all possible room candidates. The feasibility of the proposed method is validated by using simulated and measured RIRs.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their constructive feedback on the manuscript and also acknowledge Youssef El Baba for helpful suggestions, Felix Knauff for his assistance in experiments and Valentin Schilling for reverberation time measurements.

REFERENCES

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

[2] F. Ribeiro, D. Ba, C. Zhang, and D. Florêncio, "Turning enemies into friends: Using reflections to improve sound source localization," in *Proc. Int. Conf. Multimedia Expo.*, 2010, pp. 731–736.

[3] P. A. Naylor and N. D. Gaubitch, *Speech Dereverberation*. Berlin, Germany: Springer, 2010.

[4] P. K. T. Wu, N. Epain, and C. Jin, "A dereverberation algorithm for spherical microphone arrays using compressed sensing techniques," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2012, pp. 4053–4056.

[5] A. Canclini, D. Markovi, F. Antonacci, A. Sarti, and S. Tubaro, "A room-compensated virtual surround system exploiting early reflections in a reverberant room," in *Proc. Eur. Signal Process. Conf.*, Aug. 2012, pp. 1029–1033.

[6] J. Scheuing and B. Yang, "Disambiguation of TDOA estimation for multiple sources in reverberant environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 1479–1489, Nov. 2008.

[7] C. M. Zannini, A. Cirillo, R. Parisi, and A. Uncini, "Improved TDOA disambiguation techniques for sound source localization in reverberant environments," in *Proc. Int. Symp. Circuits Syst.*, 2010, pp. 2666–2669.

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

[9] P. Annibale, J. Filos, P. A. Naylor, and R. Rabenstein, "Geometric inference of the room geometry under temperature variations," in *Proc. Int. Symp. Control, Commmun. Signal Process*, 2012, pp. 1–4.

[10] J. Filos, E. A. Habets, and P. A. Naylor, "A two-step approach to blindly infer room geometries," in *Proc. Int. Workshop Acoust. Echo Noise Control*, 2010, pp. 1–4.

[11] F. Antonacci *et al.*, "Inference of room geometry from acoustic impulse responses," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 10, pp. 2683–2695, Dec. 2012.

[12] J. Filos, A. Canclini, F. Antonacci, A. Sarti, and P. A. Naylor, "Localization of planar acoustic reflectors from the combination of linear estimates," in *Proc. Eur. Signal Process. Conf.*, 2012, pp. 1019–1023.

[13] L. Remaggi, P. J. B. Jackson, W. Wang, and J. A. Chambers, "A 3D model for room boundary estimation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2015, pp. 514–518.

[14] L. Remaggi, P. J. B. Jackson, P. Coleman, and W. Wang, "Acoustic reflector localization: Novel image source reversion and direct localization methods," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 2, pp. 296–309, Feb. 2017.

[15] I. Dokmanić, R. Parhizkar, A. Walther, Y. M. Lu, and M. Vetterli, "Acoustic echoes reveal room shape," *Proc. Nat. Acad. Sci.*, vol. 110, no. 30, pp. 12 186–12 191, 2013.

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

[17] N. D. Gaubitch, W. B. Kleijn, and R. Heusdens, "Auto-localization in ad-hoc microphone arrays," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 106–110.

[18] A. M. Torres, J. J. Lopez, B. Pueo, and M. Cobos, "Room acoustics analysis using circular arrays: An experimental study based on sound field plane-wave decomposition," *J. Acoust. Soc. Amer.*, vol. 133, no. 4, pp. 2146–2156, 2013.

[19] H. Kim, L. Remaggi, P. J. Jackson, F. M. Fazi, and A. Hilton, "3D room geometry reconstruction using audio-visual sensors," in *Proc. Int. Conf. 3D Vision*, Oct. 2017, pp. 621–629.

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

[21] Y. El Baba, A. Walther, and E. A. P. Habets, "Time of arrival disambiguation using the linear Radon transform," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2017, pp. 106–110.

[22] Y. El Baba, A. Walther, and E. A. Habets, "3D room geometry inference based on room impulse response stacks," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 5, pp. 857–872, May 2018.

[23] E. Mabande, H. Sun, K. Kowalczyk, and W. Kellermann, "On 2d localization of reflectors using robust beamforming techniques," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2011, pp. 153–156.

[24] L. E. Kinsler, A. R. Frey, A. B. Coppens, and J. V. Sanders, *Fundamentals of Acoustics*, 4th ed. Hoboken, NJ, USA: Wiley, 1999.

[25] F. Ribeiro, D. Florêncio, D. Bai, and S. Zhang, "Geometrically constrained room modeling with compact microphone arrays," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 5, pp. 1449–1460, Jul. 2012.

[26] M. F. S. Galvez, D. Menzies, and F. M. Fazi, "Dynamic audio reproduction with linear loudspeaker arrays," *J. Audio Eng. Soc.*, vol. 67, no. 4, pp. 190–200, 2019.

[27] M. Foco, P. Polotti, A. Sarti, and S. Tubaro, "Sound spatialization based on fast beam tracing in the dual space," in *Proc. Int. Conf. Digit. Audio Effects (DAFx-03)*, Sep. 2003, pp. 1–5, Paper 52.

[28] F. Antonacci, M. Foco, A. Sarti, and S. Tubaro, "Fast tracing of acoustic beams and paths through visibility lookup," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 4, pp. 812–824, May 2008.

[29] D. Markovic, A. Canclini, F. Antonacci, A. Sarti, and S. Tubaro, "Visibility-based beam tracing for soundfield rendering," in *Proc. Int. Workshop Multimedia Signal Process.*, 2010, pp. 40–45.

[30] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. R. Stat. Soc. Ser. B Methodol.*, vol. 58, no. 1, pp. 267–288, 1996.

[31] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. R. Stat. Soc. Ser. B Methodol.*, vol. 67, no. 2, pp. 301–320, 2005.

[32] P. M. Peterson, "Simulating the response of multiple microphones to a single acoustic source in a reverberant room," *J. Acoust. Soc. Amer.*, vol. 80, no. 5, pp. 1527–1529, 1986.

[33] E. A. P. Habets, "Single- and multi-microphone speech dereverberation using spectral enhancement," Ph.D. dissertation, Dept. Elect. Eng., Eindhoven Univ. Technol., Eindhoven, The Netherlands, 2007.

[34] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *J. Stat. Softw.*, vol. 33, no. 1, pp. 1–22, 2010.

[35] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. 2nd ed., Cambridge, U.K.: Cambridge Univ. Press, 2003.

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

[37] ISO 3382-2:2008(EN), "Acoustics – Measurement of room acoustic parameters – Part 2: Reverberation time in ordinary rooms," International Organization for Standardization, Geneva, CH, Standard, 2008.

[38] J. A. Fessler and L. Rogers, "Spatial resolution properties of penalized-likelihood image reconstruction: Space-invariant tomographs," *IEEE Trans. Image Process.*, vol. 5, no. 9, pp. 1346–1358, Sep. 1996.

[39] A. Farina, "Advancements in impulse response measurements by sine sweeps," in *Proc. Audio Eng. Soc. Conv.*, May 2007, pp. 1–21, Paper 7121.

Cagdas Tuna received the B.Sc. degrees in telecommunications engineering and electronics engineering (double major program) from Istanbul Technical University, Istanbul, Turkey, in 2006 and 2007, respectively, and the M.Sc. (as a Fulbright scholar) and Ph.D. degrees in electrical and computer engineering from the University of Illinois at Urbana-Champaign, Champaign, IL, USA, in 2009 and 2014, respectively. From 2014 to 2017, he was a Postdoctoral Researcher with the University of Illinois' Advanced Digital Sciences Center (ADSC), Singapore. Since 2017, he has been a Research Associate with the International Audio Laboratories Erlangen (a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg and Fraunhofer IIS, Erlangen, Germany). His research interests include spatial audio signal processing, room acoustics, sensor-array signal processing and inverse problems.

Antonio Canclini was born in Sondalo, Italy, on December 19, 1983. He received the M.Sc. degree (*cum laude*) in computer science and engineering and the Ph.D. degree (*cum laude*) in information technology from the Politecnico di Milano, Milan, Italy, in 2008 and 2012, respectively. He is currently a Contract Professor with the Politecnico di Milano, where he is also a Postdoctoral Researcher with the Image and Sound Processing Group, Dipartimento di Elettronica, Informazione e Bioingegneria. His research interests focus on space-time audio signal processing (localization of acoustic sources and reflectors, spatial sound field rendering).

Federico Borra (Student Member, IEEE) received the B.Sc. and M.Sc. degrees (*cum laude*) in computer engineering from the Politecnico di Milano, Milan, Italy, in 2014 and 2016, respectively, and the Ph.D. degree in information engineering in 2020 from the Politecnico di Milano, Milan, Italy, where he is currently a Postdoctoral Researcher. His main research interests concern space-time audio signal processing.

Philipp Götz received the Bachelor of Arts degree from the Department of Sonology, Royal Conservatory in The Hague, The Hague, The Netherlands, in 2010, and the M.Sc. degree in audio communication from the Technical University in Berlin, Berlin, Germany, in 2015. Since 2012, he has been working with the Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany, first as a Student Intern, and, starting from 2015, as a Research Associate. His research interests include acoustic signal processing and spatial sound field analysis.

Fabio Antonacci (Member, IEEE) was born in Bari, Italy, on July 26, 1979. He received the Laurea degree in telecommunication engineering and the Ph.D. degree in information engineering from the Politecnico di Milano, Milan, Italy, in 2004 and 2008, respectively. He is currently an Assistant Professor with the Politecnico di Milano. His research interests focus on space-time processing of audio signals, for both speaker and microphone arrays (source localization, acoustic scene analysis, rendering of spatial sound) and on modeling of acoustic propagation. He is a member of the IEEE Audio and Acoustic Signal Processing Technical Committee and of the EURASIP SAT on Audio, Speech and Music Signal Processing.

Andreas Walther received the Dipl. Ing. degree in media technology from the Technical University of Ilmenau, Ilmenau, Germany, in 2005 and the Ph.D. degree from the Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland, in 2013, for his work on perception and reproduction of auditory spatial impression. From 2005 to 2009, he worked with the Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany, as a Research and Development Engineer. In 2009, he joined the Audiovisual Communication Laboratory, EPFL. In 2013, he joined the Semantic Audio Processing Department, Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany, and is currently the Head of the Semantic Spatial Audio Rendering group. His main research interests include spatial sound reproduction, spatial audio signal processing, room acoustics, and auditory perception.

Augusto Sarti (Senior Member, IEEE) received the Ph.D. degree in information engineering from the University of Padova, Padua, Italy, in 1993, with a joint graduate program with the University of California, Berkeley, Berkeley, CA, USA. In 1993, he joined the Faculty of the Politecnico di Milano, Milan, Italy, where he is currently a Full Professor. In 2013, he also joined the University of California, Davis. He coordinates the activities of the Musical Acoustics Laboratory and the Sound and Music Computing Laboratory of the Politecnico di Milano. He promoted/coordinated and/or contributed to numerous European projects in the area of multimedia signal processing. He has coauthored more than 300 scientific publications on international journals and congresses and numerous patents in the multimedia signal processing area. His main research interests are in the area of audio and acoustic signal processing, with particular focus on sound analysis, synthesis, and processing; space-time audio processing; geometrical acoustics; music information extraction and music modeling. He served in the IEEE Technical Committee on Audio and Acoustics Signal Processing for two terms. He was an Associate Editor for the IEEE/ACM, Transactions on Audio, Speech, and Language Processing, and as a Senior Area Editor for the IEEE Signal Processing Letters, and in 2017 he received the "Outstanding Editorial Board Member Awards" by the IEEE Signal Processing Society. He co-chaired the IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS-05); he chaired the Digital Audio Effects conference (DAFx-09); and he co-chaired the IEEE International Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA-19). He was in the organizing committees of numerous International Conferences, including IEEE ICASSP-14, the ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC-09), the IEEE International Workshop on Haptic Audio-Visual Environment and Game (HAVE-09), and the European Signal Processing Conference (EUSIPCO-2018). He is currently serving in the EURASIP board of directors.

Emanuël A. P. Habets (Senior Member, IEEE) received the B.Sc. degree in electrical engineering from the Hogeschool Limburg, The Netherlands, in 1999, and the M.Sc. and Ph.D. degrees in electrical engineering from the Technische Universiteit Eindhoven, Eindhoven, The Netherlands, in 2002 and 2007, respectively.

He is an Associate Professor with the International Audio Laboratories Erlangen (a joint institution of the Friedrich-Alexander-Universitat Erlangen-Nürnberg and Fraunhofer IIS), and the Head of the Spatial Audio Research Group, Fraunhofer IIS, German. From 2007 to 2009, he was a Postdoctoral Fellow with the Technion - Israel Institute of Technology and at the Bar-Ilan University, Israel. From 2009 to 2010, he was a Research Fellow with the Communication and Signal Processing Group, Imperial College London, U.K. His research activities center around audio and acoustic signal processing, and include spatial audio signal processing, spatial sound recording and reproduction, speech enhancement (dereverberation, noise reduction, echo reduction), and sound localization and tracking.

Dr. Habets was a member of the organization committee of the 2005 International Workshop on Acoustic Echo and Noise Control (IWAENC) in Eindhoven, The Netherlands, a General Co-Chair of the 2013 International Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) in New Paltz, New York, and a General Co-Chair of the 2014 International Conference on Spatial Audio (ICSA), Erlangen, Germany. He was a member of the IEEE Signal Processing Society Standing Committee on Industry Digital Signal Processing Technology (2013–2015), a Guest Editor for the IEEE Journal of Selected Topics in Signal Processing and the *EURASIP Journal on Advances in Signal Processing*, an Associate Editor for the IEEE Signal Processing Letters (2013–2017), and the Editor in Chief for the *EURASIP Journal on Audio, Speech, and Music Processing* (2016–2018). He was the recipient, with S. Gannot and I. Cohen, of the 2014 IEEE Signal Processing Letters Best Paper Award. He is currently a member of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing, a member of the EURASIP Technical Activities Board, and Chair of the EURASIP Technical Area Committee on Acoustic, Speech and Music Signal Processing.