3D Room Geometry Inference Based on Room Impulse Response Stacks

Youssef El Baba, *Student Member, IEEE*, Andreas Walther, and Emanuël A. P. Habets, *Senior Member, IEEE*

***Abstract*—Room geometry inference is concerned with the localization of reflective boundaries in an enclosed space. This paper outlines a method for inferring room geometry based on the positions of loudspeakers and real or image microphones, which are computed using sets of times of arrival (TOAs) obtained from room impulse responses (RIRs). These RIRs describe the acoustic propagation between the loudspeakers in an array and a single microphone. First, peaks corresponding to TOAs in these RIRs are detected and labeled using an automated method. Second, the labeled TOA sets are used to estimate the real and image microphone positions, with knowledge of the loudspeaker array geometry. Third, using all these positions, the positions of reflection points on the available reflectors in the room are determined. The reflection points determine the reflectors' locations and orientations. This approach is largely automated and usable in real-world scenarios.**

***Index Terms*—Image model, TOA disambiguation, echo labeling, reflection point localization, reflector localization, room geometry inference.**

I. Introduction

THE task of room geometry inference (RGI) is concerned with the localization of reflective boundaries in an enclosed space, and is of interest in several applications: 3D sound analysis and reproduction [1]–[4], robust sound source localization (SSL) [5], speaker tracking [6] and dereverberation [7]. RGI is commonly equated with reflector localization (RL) applied to all physical walls in a given room. RL methods use times of arrivals (TOAs) of the direct-path and reflections — peaks in room impulse responses (RIRs) from different microphone and loudspeaker position combinations — to infer the locations and orientations of planar reflectors. First-order reflections are especially of interest because they characterize physical walls bounding the room. The largest family of RL methods relies on ellipse geometry [8]–[23]; however, the ellipse-based problem formulation has a dual formulation [24] that relies on hyperbola geometry [25], [26]. Other methods rely on beamforming [27], [28] or directivity-based schemes [29], and the remainder [5], [30]–[41] rely on various other schemes. For RL, TOAs need to be separated into sets, each set belonging to a single reflector [25]. These sets are used individually with the measurement position, either known or estimated using SSL, to define multiple constraints which together localize a reflector.

The ellipse-based RGI family [8]–[23] operates according to the approach presented for 2D RL in [8]; this method is equivalent to an inversion of the first-order image model [42]. Multiple ellipses between pairs of microphones and loudspeakers are inferred from the TOAs, and a reflector is determined as a common tangent to these ellipses. The work in [14], [15] tests the effect of temperature variations on the approach, whereas the work in [16], [17] enhances the performance of the same approach by circumventing the non-linearity of the RL estimator's cost function, thereby simplifying the optimization of the cost function and refining it. Additionally, the reflection point localization scheme presented in [22] equally relies on ellipse geometry. For extension to 3D, the authors in [13] switch to ellipsoids, whereas those in [12] divide the 3D space into 2D sub-planes. The switch to ellipsoids is also described in [18]–[21]; these four contributions also enhance robustness to noise and use the random sample consensus algorithm [43] to search for the reflectors [21]. More generally, the work in [23] extends [8] to 3D simultaneous localization and mapping.

Alternative RGI methods have few common characteristics with each other or with the ellipse-based family. The method in [30] localizes reflections in an auditorium by estimating time differences of arrivals (TDOAs), i.e., differences between TOAs, using cross power spectrum phase. Other approaches employ beamforming techniques, such as steered response power [27], [28], and [35] similarly uses likelihood maps. The method in [29], [31], [32] analyzes sound intensity vectors for reflection point localization. Furthermore, in a considerably different approach, the authors in [5], [33], [34] compare measured RIRs to an RIR database for fitting a shoebox room model. A more recent method [36] employs Euclidean distance matrices for RL. Follow-up publications extend [36] to simultaneous localization and mapping [39], [44], self-calibration [40], and membrane shape detection of non-planar reflective surfaces [37], [41]. Alternatively, RL is achieved in [38] by analyzing the resonant frequencies between a source and a receiver aligned orthogonally between two reflectors.

Manuscript received June 21, 2017; revised October 16, 2017 and November 30, 2017; accepted November 30, 2017. Date of publication December 15, 2017; date of current version March 15, 2018. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Roland Badeau. *(Corresponding author: Youssef El Baba.)*

Y. El Baba and E. A. P. Habets are with the International Audio Laboratories Erlangen (A joint institution of the Friedrich-Alexander-University Erlangen-Nürnberg and Fraunhofer IIS), Erlangen 91058, Germany (e-mail: youssef.elbaba@audiolabs-erlangen.de; emanuel.habets@audiolabs-erlangen.de).

A. Walther is with the Fraunhofer Institute for Integrated Circuits, Erlangen 91058, Germany (e-mail: andreas.walther@iis.fraunhofer.de).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2017.2784298

Topics related to RGI include SSL and echo labeling. SSL refers to estimating the locations of sources using ranges (via TOAs) or range differences (via TDOAs), whereas echo labeling refers to determining which TOAs or TDOAs correspond to the same reflector. More advanced RGI-related topics include 3D room imaging [45], RGI using only a single RIR [26], [46]–[48] and full RGI automation [49]. Echo labeling, referred to as disambiguation in the context of RGI, is of particular interest because it has not been extensively dealt with in the RGI literature, with few exceptions [9], [11], [36], [50], [51], although it is a critical component. For instance, some RGI authors use compact arrays to effectively avoid labeling TOAs [5], [21], [33], [34].

The presence of multiple reflectors in a room produces multiple and potentially overlapping reflections. This makes labeling echoes more involved. Such echo labeling is needed because RL methods give phantom results if provided with sets of echoes belonging to different reflectors [25]. Some graph-theory-based methods [25], [52], [53] disambiguate TDOAs, not TOAs, which makes them vulnerable to tight margins for TDOA error.[1] The method in [25] has particularly been relied on by RGI authors, but it uses a combinatorial scheme involving TDOAs. Other methods [36], [50], [51], [54] use combinatorial schemes directly involving TOAs instead of TDOAs. All these methods [25], [36], [50]–[54] result in a high computational complexity, due to the impractically high number of TDOA or TOA combinations, with [51] being the most efficient due to substantial search reduction. Disambiguation is also tackled in [10], [11] via the generalized Hough transform built on top of [8]. This more promising technique is further developed in [9], but the source is moved sequentially next to each reflector, which provides a good initial starting solution that is later refined. The present authors' previous work [55] also addresses the echo labeling problem using image-processing techniques. In contrast to [9], neither the echo-labeling scheme in [55] nor the scheme in the present paper requires moving the source sequentially next to each reflector.

Methods for SSL apply concepts initially developed in the context of satellite localization to the acoustic context. The method in [56] is the most used among the four similar TDOA-based methods in [56]–[59], due to its simple and efficient closed-form least-squares formulation. A concise formulation and overview of these SSL methods is available in [60]. The more recent method in [61] performs SSL using TOAs instead of TDOAs; this is done via a non-optimal least-squares-based estimator, or alternatively via a version of it improved with regularization. In [62], [63], SSL performance was enhanced by exploiting the common spatial coherence of RIRs; this was based on prior knowledge about the room geometry. In the same vein, [5], [34] invert the RGI problem by assisting SSL with the estimated room geometry. Similarly, the authors in [64] use image microphones for making SSL more robust.

Most RL methods assume 2D use cases and use circular, planar or randomly-distributed [36] microphone arrays accompanied by a single omnidirectional loudspeaker. Similarly, 3D RL methods preponderantly use spherical microphone arrays. In all cases, an ideal transducer setup exhibiting a sufficiently high spatial diversity is assumed. Accordingly, for a loudspeaker at a fixed position in a shoebox room, a 2D or 3D microphone array is used in the 2D or 3D case to localize up to four or six reflectors, respectively. In our own and other applications, arrays with such high spatial diversity are rarely available, introducing a significant geometrical ambiguity. For instance, a 2D loudspeaker array is missing one dimension for 3D RL; hence, using it for 3D RL introduces a front/back ambiguity because of the reduced spatial diversity, makes the problem ill-posed and prevents the direct use of existing RL methods. Ill-posedness is especially noticeable when i) SSL estimates become severely inaccurate, and ii) RL provides erroneous yet plausible solutions.

In this contribution, and in contrast to previous works, we consider precisely such a 2D rectangular loudspeaker[2] array with reduced spatial diversity accompanied by a single omnidirectional microphone for 3D RGI. Such rectangular loudspeakers arrays can find application in future reproduction systems that place multiple loudspeakers around video screens (e.g., [3]).

In previous RGI algorithms, SSL is only used to spatio-temporally synchronize the system by estimating the location of the real microphone or source, using the TDOAs derived from the peaks of the direct sound wavefront in the RIRs. In contrast, we use SSL to estimate the locations of both real and image microphones. For this, we assume a synchronized[3] setup to directly use TOAs instead of TDOAs, thereby increasing the margins for error considerably. Our array is densely populated with loudspeaker drivers, and hence contains a larger number of transducers than commonly found in the literature;[4] in addition, it exhibits a larger aperture. Our method benefits from these two last properties, albeit not equally. More specifically, the dense transducer placement is beneficial, whereas the large number of transducers is less beneficial. In contrast, for strictly planar reflectors, the large aperture is detrimental to the method. The crucial property is the linearity of the four array branches; we rely on this linearity for echo labeling.[5] More generally, the array does not need to be rectangular, it suffices for it to be planar; nonetheless, this planarity introduces the aforementioned drawback of a geometrical ambiguity in 3D.

The RGI method outlined in this work consists of the following steps. First, peaks corresponding to TOAs in the RIRs are

[1] For example: with 48 kHz sampling, a TDOA of a wavefront between sensors 40 cm apart and a source 2.6 m away takes a value of 4.46 samples.

[2] This setup involves a change of perspective, namely in using a loudspeaker array instead of a microphone array, but this change is straightforward due to acoustic reciprocity [24].

[3] More specifically, we assume that all transducers use the same sampling frequency, and that any latency in measurements is known.

[4] It should be noted that other authors did use large numbers of transducers, especially when employing beamforming techniques. One example is the work in [27], which used 60 microphones. Moving the array next to each wall and repeating the measurements [9], [13] is also equivalent to using a large number of transducers. Also note that the use of better transducer arrays is justified in light of the method being able to perform disambiguation automatically, as opposed to with supervised or highly combinatorial schemes as in most of the RGI literature.

[5] The scheme we employ currently only works for linear arrays; however, extensions of this scheme to non-linear arrays are theoretically possible.

Fig. 1. Block diagram of the proposed IMRL algorithm. Boxes bounded by dashed lines denote components that currently require supervision.

detected and labeled using an extended version of the method in [55], in which the labeling is limited to 2D and a single linear array. The extension to 3D, as well as the development of a graph-based method to associate labels from multiple linear arrays, are the main contributions of this paper. Second, the labeled TOA sets are used to estimate the real and image microphone positions, with knowledge of the loudspeaker array geometry. Third, using the estimated microphone positions and the array geometry, the positions of reflection points on the available reflectors are determined, using the method in [22]. Finally, the reflection points determine the reflectors' locations and orientations.

The remainder of this paper is organized as follows: Section II formulates the problem and gives an overview of the proposed solution, Section III more specifically explains how TOAs are detected and disambiguated, Section IV details the adopted SSL scheme, Section V details the reflection point localization and RL schemes, Section VI evaluates the performance of the proposed solution and conclusions are given in Section VII.

II. Problem Formulation and Proposed Solution

A. Notation and Problem Formulation

Throughout this paper, scalars are denoted by non-bold, italic letters as in m and M. Vectors are denoted by lowercase, bold, non-italic letters as in $\mathbf{m}$, and are also used to describe 2D and 3D positions, with the scalar m_i referring to the i-th element or coordinate; matrices are indicated by uppercase, bold, non-italic letters as in $\mathbf{M}$, with $\mathbf{M}[j,i]$ referring to the element at the j-th row and i-th column. Uppercase, non-bold, non-italic letters as in M indicate general objects and geometrical entities such as points, lines and circles; sets are indicated by uppercase, calligraphic letters as in $\mathcal{M}$. The particular index r refers to one of R real or image reflectors. Additionally, the notations $\widehat{\cdot}$ and $\bar{\cdot}$ refer to estimated and interim entities, respectively.

Given an array of L loudspeakers with a single omnidirectional microphone, and assuming that the acoustic propagation can be modeled by a linear time-invariant filter, the RIR of the filter between the j-th loudspeaker and the microphone (notwithstanding noise) can be expressed by

$$h_j(t) = \alpha_{0j}\delta(t-\tau_{0j}) + \sum_{r=1}^{R} \alpha_{rj}\delta(t-\tau_{rj}) + \eta_j(t)\,,$$

where α_{0j} and α_{rj} are the attenuation coefficients of the direct and reflection paths, respectively. The function $\delta(t)$ represents the delta function, $\eta_j(t)$ represents possible measurement noise and t denotes time. The TOAs τ_{0j} that arrive from the real microphone to the L loudspeakers and the TOAs τ_{rj} $(r \in \{1..R\})$ that arrive from the image microphones correspond to the direct and reflected wavefronts, respectively; they form the sets

$$\mathcal{T}_r = \left\{\tau_{rj} : \forall j \in \{1..L\}\right\}. \quad (1)$$

These RIRs and the known relative loudspeaker positions in the array, i.e., the array geometry, constitute the input data. TOAs need to be detected and disambiguated into separate sets $\{\mathcal{T}_r : \forall r \in \{0\,..\,R\}\}$. The aim is to obtain from these TOA sets the desired plane equations $\langle \mathbf{n}_r, \mathbf{x}\rangle + o_r = 0$ characterizing the different reflectors' planes, where $\langle\,.\,,\,.\,\rangle$ denotes the scalar product between vectors, $\mathbf{n}_r$ and o_r denote the r-th plane's normal vector and offset, and $\mathbf{x}$ denotes the general 3D coordinate vector.

B. Overview of the Proposed Method

The proposed image microphone reflector localization (IMRL) algorithm uses the input data to obtain the reflectors' equations in three main steps, as depicted in Fig. 1.

A *convenient* spatio-temporal representation of RIRs is used, referred to as an RIR stack; it is a vertical stacking of RIRs as rows in an image, where amplitudes are translated into brightness values or color (Fig. 2). For RIRs measured from the loudspeakers of a uniform loudspeaker array to a single microphone, an RIR stack is a uniform sampling of sound propagation in the room in space and time.

Once RIRs are visualized in stacks, distinct structures appear, which correspond to acoustic wavefronts. These arrive at the microphone from the sources and their images mirrored w.r.t. the reflectors in the room [42]. Within each linear branch — referred to as a sub-array — of the rectangular loudspeaker array, these structures can be approximated by straight lines on the stack — referred to as stack-lines (SLs) — especially when their corresponding real or image microphone lies far from the loudspeaker array, and when the sub-array's aperture is small. In terms of TOA disambiguation, each SL corresponds to a TOA set disambiguated within a single sub-array. Therefore, SLs characterize the wavefronts within a single sub-array unambiguously. This is because from the perspective of a single microphone, the direct and reflected wavefronts from a linear sub-array can be reasonably considered planar wavefronts for common room sizes, array sizes and transducer configurations.

As a first step in IMRL, SLs are initially detected (Fig. 4) within each of the sub-arrays' portions of the RIR stack using image processing techniques [55] such as the linear Radon transform. The linear Radon transform is similar to beamforming when applied in the acoustic context, but its original purpose in image processing is line detection. Using this initial step it is possible to disambiguate TOAs on RIR stacks measured with linear transducer arrays but not with other array geometries; hence, a further step associates SLs across sub-arrays, since independent SLs in separate linear arrays do not offer enough spatial diversity for 3D RL. The resulting SL groups, consisting of SLs from multiple linear transducer arrays (Fig. 7), are then used to construct full TOA sets $\{\mathcal{T}_r : \forall r \in \{0\,..\,R\}\}$ (Fig. 9) for 3D RL, each set $\mathcal{T}_r$ belonging to a single real or image

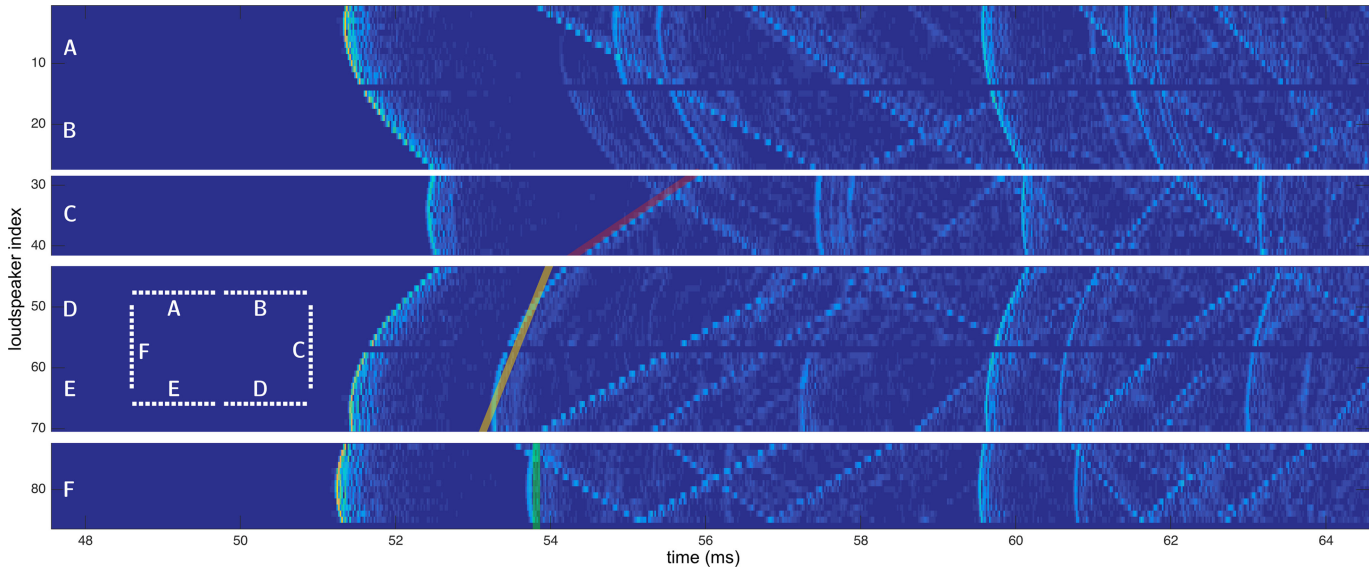

Fig. 2. Example RIR stack from real measurements. The white overlay figure depicts the actual loudspeaker array used for measurement, with the capital letters A, B, C, D, E and F indicating the individual uniform linear arrays used in the measurement; A and B form a single sub-array, similarly D and E. The blue gaps in the center rows of the first and third sub-arrays correspond to measurement gaps between A and B, and D and E, respectively. Each individual loudspeaker contributes a row in one of the individual sub-arrays' RIR stack portions (separated by white spaces), with yellow indicating high absolute amplitudes and blue indicating low absolute amplitudes. Red, yellow and green highlights correspond to same-color circles in Fig. 6 (Section III-B).

microphone. These steps correspond to TOA detection and disambiguation (Section III).

Second, each disambiguated set, with its TOAs defining spatial ranges and with knowledge of the array geometry, is used to estimate the position of a real or image microphone using SSL (Section IV). Although SSL techniques are traditionally used for localizing sound *sources*, they can nonetheless be used for localizing sound *receivers*, thanks to acoustic reciprocity. In IMRL, this second step alone is sufficient for tackling the front/back ambiguity in our setup; this is achieved by assuming that the array is positioned[6] near and approximately-parallel to one of the reflectors. Since the real and most image microphone positions (except those associated with the nearby wall) lie in the region in front of the array, it is sufficient for all practical purposes to assume most positions have a positive x coordinate (more details in Section IV).

Third, once these positions are retrieved, they are used in a projective scheme to locate the positions of reflection points on the reflectors of the room (Fig. 10). In the absence of SSL or numerical errors, the points perfectly align on the reflectors; they accordingly determine the reflectors' positions and orientations (Section V).

This procedure estimates physical, first-order reflectors corresponding to the boundaries of the room when provided with first-order reflections. But when provided with second- or higher-order reflections, it estimates nonexistent reflectors. If physical room geometry reconstruction is sought, additional processing is needed to select and retain only the physical reflectors, using for instance a bounded-box logic [8]. This can be equivalently achieved either by distinguishing and retaining first-order from higher-order TOA sets, or first-order from higher-order image microphones. In this paper, this reflector selection is supervised.

[6]This and the echo-labeling scheme are the only places in our algorithm where this placement is exploited.

III. TOA Detection and Disambiguation

The scheme described in this section jointly detects and disambiguates TOAs via SL detection followed by SL association; It alleviates the need for TOA detection in individual RIRs.

To remove irrelevant data, the temporal length of the RIRs is cut to a predetermined truncation time [55]. The RIR stack is then constructed as a matrix $\bar{\mathbf{R}}$ with $\bar{\mathbf{R}}[j,i] = h_j(i)$, where $j \in \{1..L\}$ is the loudspeaker index and i is the sample index. Each RIR sample corresponds to a pixel on the RIR stack.

Since acoustic wavefronts produce both positive and negative amplitude deviations around zero, we use this to further enhance $\bar{\mathbf{R}}$. In particular, the location and sign of the peak with the highest absolute amplitude across the stack are determined, and all RIR amplitudes with an opposite sign are set to zero. This processing retains only one side of the RIR waveforms: the positive side or negative side depending on the polarity of the aforementioned peak; this scheme is a digital variant of electrical half-wave rectification, complemented by polarity detection. The aim is to remove redundant lobes from the loudspeakers' impulse responses, as these spurious lobes could be mistaken for genuine TOAs. The resulting stack $\mathbf{R}$ is given by

$$\mathbf{R} = \left| \mathbf{I}\left(\bar{\mathbf{R}}\left[\underset{j,i}{\operatorname{argmax}} \left|\bar{\mathbf{R}}[j,i]\right|\right], \bar{\mathbf{R}}\right) \circ \bar{\mathbf{R}} \right|,$$

$$\mathbf{I}(u,\mathbf{M})[j,i] = \begin{cases} 1 & \text{if } \operatorname{sgn}\big(\mathbf{M}[j,i]\big) = \operatorname{sgn}(u) \\ 0 & \text{otherwise} \end{cases},$$

where $\circ$, $|.|$ and sgn denote the element-wise matrix multiplication (Hadamard product), the element-wise absolute value and the sign function, respectively.

In the following, further notations are employed: the particular index a and the capital letter A denote the index and number of sub-arrays in the rectangular loudspeaker array, respectively; $\mathcal{L}_a$ denotes the set of all loudspeakers belonging to the a-th sub-array with $l_a = |\mathcal{L}_a|$ denoting its cardinality; the position of the j-th loudspeaker is indicated by $\mathbf{s}_j$.

A. Stack-Line Detection With the Linear Radon Transform

Within each sub-array's portion of $\mathbf{R}$ (Fig. 2), the scheme in [55] is used to detect SLs in multiple steps. First, the portion is resampled to reference spatial and temporal resolutions. Second, a linear Radon transform[7] is applied to the resampled portion. Third, image processing is applied on the linear Radon transform's response. Fourth, salient peaks in the response are detected and mapped back to SLs on the original RIR stack portion. A detailed account of the these steps is given in [55]. The linear Radon transform can be understood on an high level as parallel sums of the stack along the vertical dimension, for many image rotation angles, with different sum contributions according to the pixels' values. Since these values originate from actual RIR amplitudes, the beamforming-like operation allows for distinguishing even faint SLs when these have the proper TOA alignment. It simultaneously mitigates noisy and non-aligned amplitude variations. Hence, a distinct SL in the stack translates to a strong peak in the transform (Fig. 3), for a certain angle and time bin. Conversely, each bin in the transform translates to a SL in the stack. While this transform optimally works on a *uniform* linear array, it can be made to work on any linear, *non-uniform* array by considering the nodes of the non-uniform array to be merely a subset of the nodes of a uniform array; another alternative is to use non-linear interpolation.

The r-th SL on the a-th sub-array is detected and translated into a TOA set

$$\mathcal{T}_{ra} = \left\{ \tau_{rj} \ : \ \forall j \in \mathcal{L}_a \right\}; \tag{2}$$

the sets for all SLs on the a-th sub-array are assembled into an auxiliary set of TOA sets $\mathcal{T}_a = \{\mathcal{T}_{ra} \ , \ r \in \{0 \,..\, \widehat{R} - 1\}\}$, with $\widehat{R}$ denoting the total number[8] of SLs detected in the sub-array; the TOAs are disambiguated within the a-th sub-array by construction. This SL detection is applied on all sub-arrays (Fig. 4) to obtain the set of sets $\mathcal{T}_a$ for all sub-arrays: $\mathcal{T} = \{\mathcal{T}_a : \forall a \ \in \{1 \,..\, A\}\}$.

B. Stack-Line Association

After the previous step, SLs are associated across sub-arrays via a graph-based approach, which assesses the geometrical compatibility between SLs. The key idea behind this processing is assessing to which degree a pair of SLs agree on a common real or image microphone position. Only pairs of SLs on different sub-arrays are considered valid combinations, i.e., pairs of SLs on the same sub-array are always incompatible.[9]

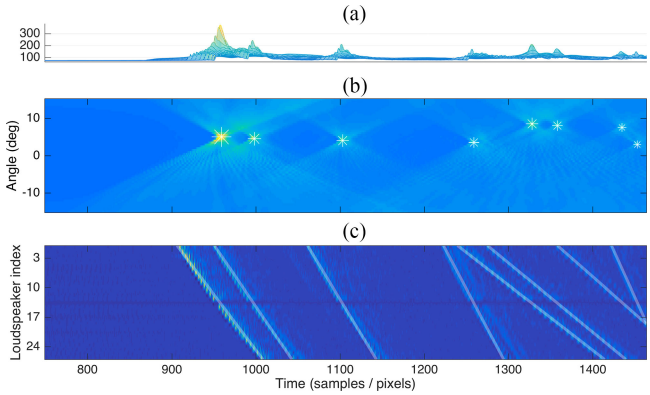

Fig. 3. A resampled RIR stack (c), its linear Radon transform (b) and the latter's side-view topography (a) [55]. Yellow encodes high values; blue encodes low values. The stack c) is obtained from real measurements, with a microphone closer to the first loudspeakers, explaining the tilt. Faint white lines on c) indicate detected SLs, and white stars on b) indicate their corresponding linear Radon transform peaks. Images b) and c) share the same horizontal (time) axis but have different vertical axes. Note that c) is from a different data set than the data set used in Fig. 2.

SL association is performed in five steps. First, a circular constraint (Fig. 5) is inferred from each SL on a specific sub-array, providing an ambiguous, circular position estimate for the real or image microphone that produced the SL. Second and third, the pairwise distances between these circles are computed and then used to construct a sparse circle-compatibility graph. Fourth, an exhaustive search finds all paths in the graph and validates corresponding circle combinations. Fifth, a conflict resolution scheme ensures that no two retrieved combinations reference the same circle. This processing detects distinct and non-overlapping cliques in the sparse compatibility graph (Fig. 7); the cliques correspond to groups of compatible SLs.

1) Circular Constraints: A SL in a specific, independent sub-array constrains the 3D location of its corresponding real or image microphone with one undetermined degree of freedom, determining a 3D circle corresponding to its locus. This is because a linear loudspeaker array can only be used to determine the direction of arrival and the range of a microphone in 2D; an ambiguity is introduced when it is employed in 3D. The construction of this circle can be understood from Fig. 5.

The r-th SL on the a-th sub-array can be characterized by its endpoints $(\tau_{r\,\mathcal{L}_a(1)} \, , \, \tau_{r\,\mathcal{L}_a(l_a)})$, corresponding to the TOAs of the first and last loudspeakers in the sub-array. These two TOAs, along with the corresponding loudspeaker positions $(\mathbf{s}_{\mathcal{L}_a(1)} \, , \mathbf{s}_{\mathcal{L}_a(l_a)})$, are abbreviated in this section with abuse of notation as τ_1, τ_2, $\mathbf{s}_1$ and $\mathbf{s}_2$, respectively. They respectively determine the ranges and centers of two TOA-range spheres, whose intersection is the circle.

[7]The Radon transform is a predecessor and more advanced variant of the Hough Transform [65].

[8]Within a sub-array, the SL detector is not guaranteed to produce exactly $R + 1$ SLs corresponding to the direct sound and R reflectors, hence the discrepancy between R and $\widehat{R} - 1$. The number $\widehat{R}$ can vary across sub-arrays.

[9]This is because a single real or image microphone produces only one SL on a single sub-array's portion of RIR stack.

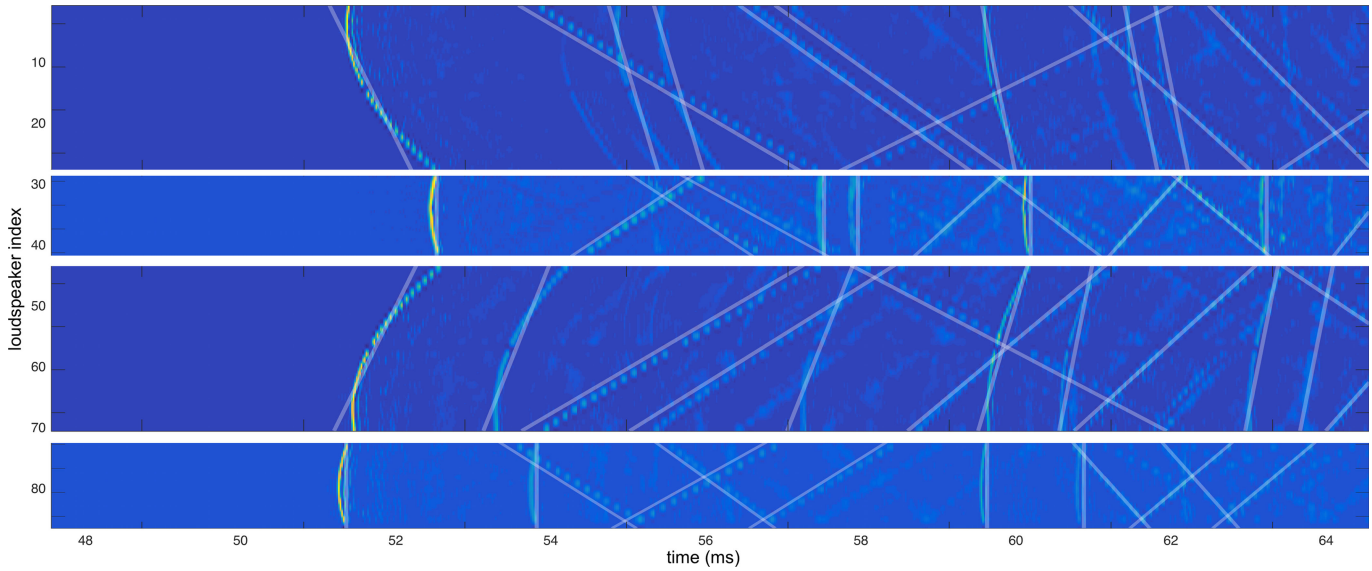

Fig. 4. SL detection with the linear Radon transform within each sub-array's portion of the RIR stack in Fig. 2 (real measurements). Yellow indicates high amplitudes and values; blue indicates low amplitudes and values. Faint white overlaid lines indicate detected SLs.

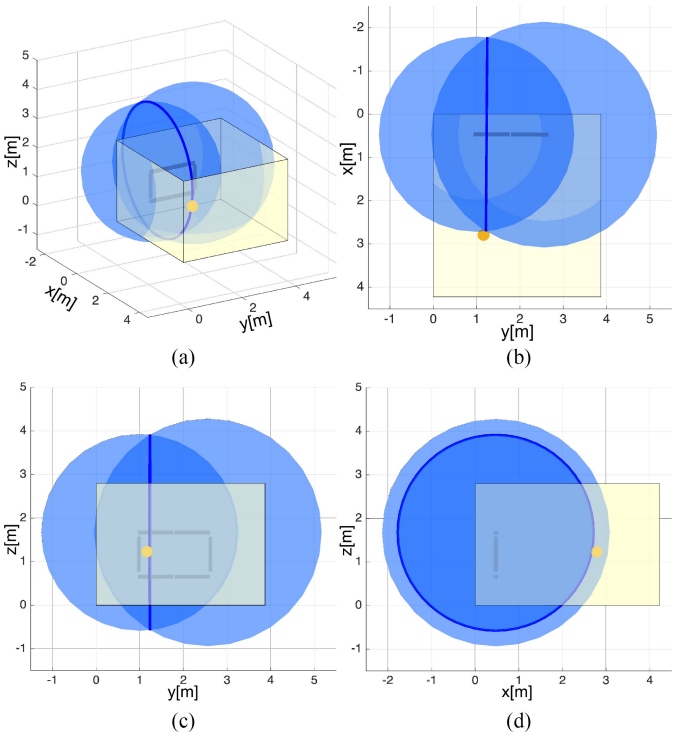

Fig. 5. Example of a circle construction for the first sub-array's RIR stack portion's ('A-B' in Fig. 4, real measurements) direct sound (earliest) SL, corresponding to the real microphone (yellow dot). Shown are the ranges estimated from the endpoints of the SL, translated into transparent blue spheres. These intersect to determine the real microphone position with one undetermined degree of freedom, hence the blue circle. The transparent beige surfaces represent the boundaries of the room, and the gray rectangle represents the loudspeaker array. (a) 3D view. (b) Top view. (c) Frontal view. (d) Side (left) view.

If the sub-array's axis going through $\mathbf{s}_1$ and $\mathbf{s}_2$ is parametrized by

$$\begin{cases} x = s_{1x} + (s_{2x} - s_{1x})p \\ y = s_{1y} + (s_{2y} - s_{1y})p \\ z = s_{1z} + (s_{2z} - s_{1z})p \end{cases}, \tag{3}$$

the position of the circle's center is obtained by taking the difference of the TOA-range spheres' equations, to determine the plane at which they intersect, and then substituting the x, y and z coordinates with their formulas in (3), to determine in turn the intersection of this plane with the sub-array's axis. This gives

$$\begin{aligned} &\|\mathbf{s}_1\|^2 - \|\mathbf{s}_2\|^2 - (c\,\tau_1)^2 + (c\,\tau_2)^2 = \\ &2\big(s_{1x} + (s_{2x} - s_{1x})p\big)(s_{1x} - s_{2x}) \\ &+ 2\big(s_{1y} + (s_{2y} - s_{1y})p\big)(s_{1y} - s_{2y}) \\ &+ 2\big(s_{1z} + (s_{2z} - s_{1z})p\big)(s_{1z} - s_{2z})\,, \end{aligned}$$

where the scalar c denotes the speed of sound and $\|.\|$ denotes the ℓ^2 norm. By solving for p we obtain

$$\begin{aligned} p_0 = &\Big(2\langle \mathbf{s}_1, \mathbf{s}_2\rangle - \|\mathbf{s}_1\|^2 - \|\mathbf{s}_2\|^2 - (c\,\tau_1)^2 + (c\,\tau_2)^2\Big) \\ &\Big/\Big(2\left(2\langle \mathbf{s}_1, \mathbf{s}_2\rangle - \|\mathbf{s}_1\|^2 - \|\mathbf{s}_2\|^2\right)\Big)\,, \end{aligned}$$

which determines the circle's center position by replacing p in (3).

The radius d_{ca} of the circle, which corresponds to the orthogonal distance between the relevant real or image microphone and the a-th sub-array, is obtained by applying Pythagoras' theorem with $d_{ca} = ((c\,\tau_1)^2 - (\|\mathbf{s}_1 - \mathbf{s}_2\|p_0)^2)^{1/2}$, where the subtrahend is the squared distance between the position of the first loudspeaker and the position of the circle's center. The 3D normal orientation of the circle is given by the sub-array's unit axis vector $\mathbf{n}_a = (\mathbf{s}_2 - \mathbf{s}_1)/\|\mathbf{s}_2 - \mathbf{s}_1\|$, because the circle and its encompassing plane are always orthogonal to the sub-array.

2) Pairwise Inter-Circle Distance Computation: The compatibility between a pair of SLs is assessed by comparing their corresponding circles: when their circles intersect in 3D, this means the SLs agree on a common real or image microphone position, in which case the SLs are compatible. The SLs of the red and yellow circles in Fig. 6 are compatible. Indeed, these circles relate to the image of the real microphone, mirrored w.r.t.

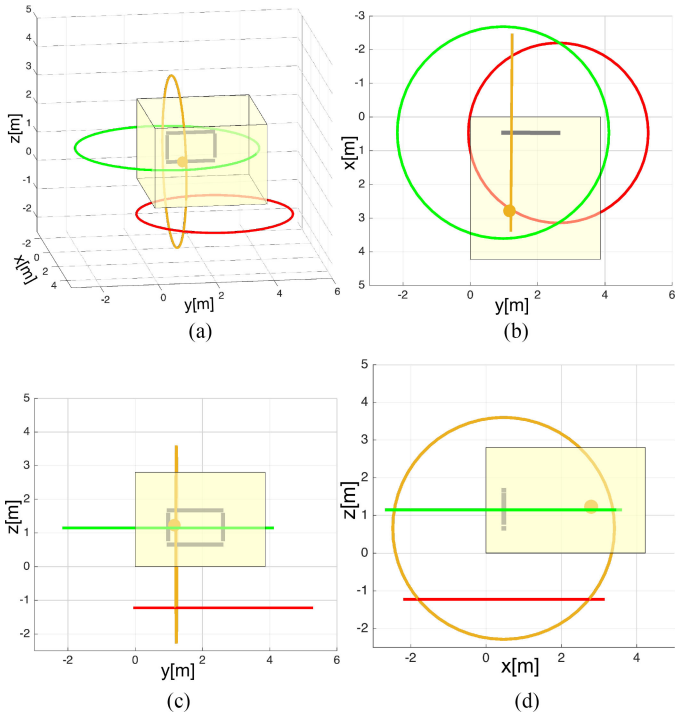

Fig. 6. Example of circles and their compatibility for SLs detected on the RIR stack in Fig. 2 (real measurements). Red, yellow and green circles correspond to SLs highlighted with the same color on sub-arrays 2, 3 and 4 of Fig. 2; for the floor, floor and left wall image microphones, respectively. The transparent beige surfaces represent the boundaries of the room, and the gray rectangle represents the loudspeaker array. (a) 3D view. (b) Top view. (c) Frontal view. (d) Side (left) view.

the floor reflector. The same is not true for the red and green circles, which relate to different image microphones. Under ideal circumstances, compatible SLs should produce such intersecting circles. However, in light of imperfect SL detection results and especially when working with real measurements — suffering from measurement as well as model error — the circles often only *nearly* intersect. The strength of the compatibility between two SLs is inversely proportional to the inter-circle distance d_{ic} between their circles. This distance can be obtained by finding the minimal distance in 3D between any two points on the two circles:

$$d_{ic}(\mathrm{C}_1, \mathrm{C}_2) = \min_{\mathbf{x}_1 \in \mathrm{C}_1\,,\, \mathbf{x}_2 \in \mathrm{C}_2} \|\mathbf{x}_1 - \mathbf{x}_2\| \,.$$

The procedure[10] to compute this distance is outlined in Appendix A, and has a lower computational complexity compared to the procedure in [66].

3) Circle-Compatibility Graph Construction: For a predetermined inter-circle distance tolerance $T_{d_{ic}}$, a pair SLs are compatible if $d_{ic} \leq T_{d_{ic}}$. Accordingly, inter-circle distances determine a sparse SL compatibility graph, an example of which is shown in Fig. 7. This graph consists of layers, corresponding to sub-arrays, and its individual vertices correspond to SLs within each sub-array.

[10]This scheme applies for any arbitrary pair of circles; however, in our application SLs on the same sub-array are always considered incompatible; we therefore do not compute the distance between their circles; instead, we restrict the distance computation to SLs lying on different sub-arrays.

The task of 3D TOA disambiguation now corresponds to detecting distinct and non-overlapping cliques in the sparse compatibility graph, in a variant of graph partitioning. In graph theory, cliques are groups of vertices where every vertex is connected to all the other vertices [67, Chapter VII, Section 34.5.1]. In general, the two graph-theory problems of clique detection and graph partitioning are NP-complete [67, Chapter VII, Section 34.5.1] and NP-hard, respectively; i.e., they belong to a set of algorithmic problems involving inherently high — usually combinatorial or exponential — computational complexity. Consequently, in the absence of a more efficient analytical, closed-form or iterative method, finding the globally optimal solution (the correct SL combinations) requires an exhaustive and/or heuristic method. The algorithm therefore exhaustively and heuristically searches for the aforementioned cliques in the sparse compatibility graph.

4) Exhaustive Path Search and Validation: Each layer in the sparse compatibility graph (Fig. 7), from first (top) to last (bottom), can be considered a measure of depth. If the graph and its edges are made directive by forcing these latter to go from higher to lower depths, e.g., from sub-array one to sub-array two, a depth-first search [67, Chapter VI, Section 22.3] approach can be used to retrieve SL groups in the graph. More precisely, we customize this approach to retrieve all possible paths going from higher to lower depths in the graph; these paths correspond to basis SL groups.

This can be achieved by customizing the standard depth-first search algorithm to start at the highest, shallow-depth vertices (e.g., nodes 1.1, 1.2 and 1.3 in Fig. 8). Then, for each start vertex, add the paths between it and all its successors as possible paths (e.g. paths with depth level 1 in Fig. 8). This is repeated recursively for all the successors while obtaining the paths starting at them and going deeper (e.g. paths with depth levels 2 and 3 in Fig. 8), and concatenating these to the shallower paths to form all relevant combinations: for a search starting at a specific shallow-depth node (labeled by a unique color in Fig. 8), depth level 1 paths are crossed with all connected depth level 2 paths to form a preliminary set of multi-depth paths, these are in turn crossed with all their possible connected lower-depth paths to form fuller paths, and so on until the deepest paths are included. In contrast to standard depth-first search, repeated vertex discovery is allowed. This explains the double-edges between certain nodes in Fig. 8, these exist for example between nodes 3.1 and 4.1 due to the former node being discovered twice: once by the orange search and once by the blue search. In addition, to ensure that all genuine groups go through the next processing step, all subsets of the retrieved paths are also retrieved, which is equivalent to combining paths of depth levels 1-2-3; 1-2, 2-3, 1-4; or 1, 2 or 3 alone; this is limited to a single search (unique color in Fig. 8). The inclusion of these subsets is logical because in some cases, a group of vertices might not form a clique, while one of its subsets might nonetheless form a clique. The vertices along the retrieved paths are then taken as basis SL groups. In this way, all connected multi-sub-array SL vertex groups are obtained, they consist of at most 1 SL vertex per sub-array and at least 2 SL vertices in total.

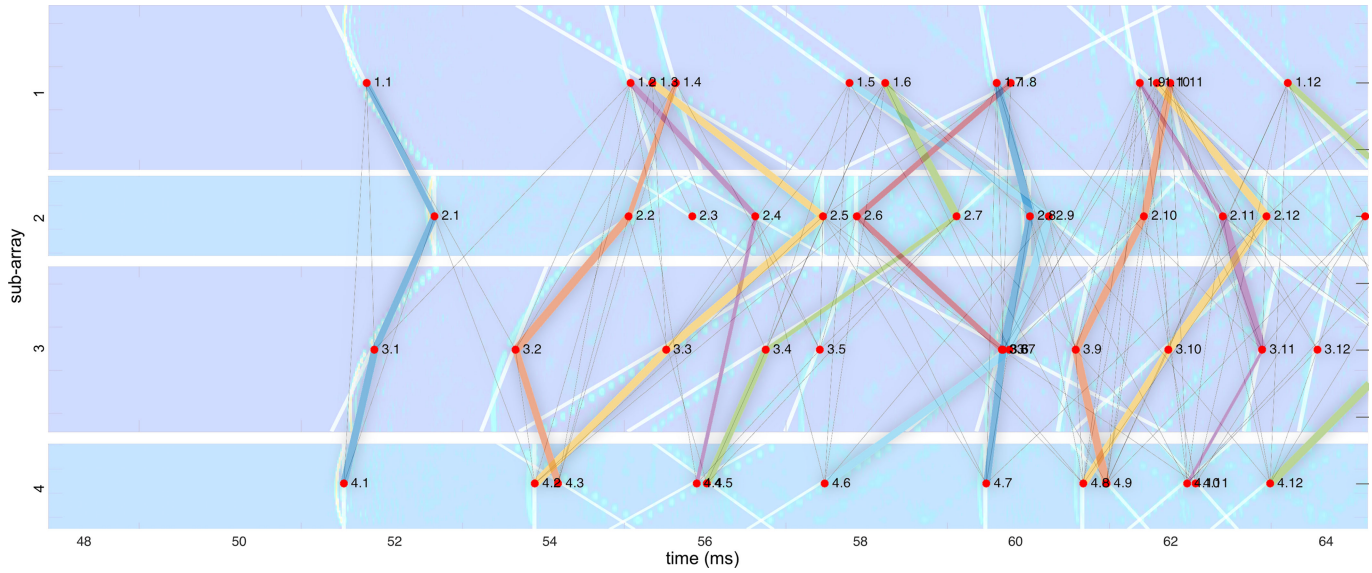

Fig. 7. SL association operating on the resulting SLs of Fig. 4 (faded background image, real measurements). Each SL in a specific, independent sub-array is considered a vertex in a sparse compatibility graph. The SLs are indicated by red dots labelled with their sub-array's index followed by their temporal order of occurrence within the sub-array's portion of the RIR stack. The association scheme associates SLs across sub-arrays according to their compatibility (indicated by thin gray edges): a valid association is one that includes at most one SL per sub-array, at least two SLs in total and in which all SLs are pairwise-compatible; forming a clique in the graph. Connections between final SL groups are distinguished by color, with repetition.

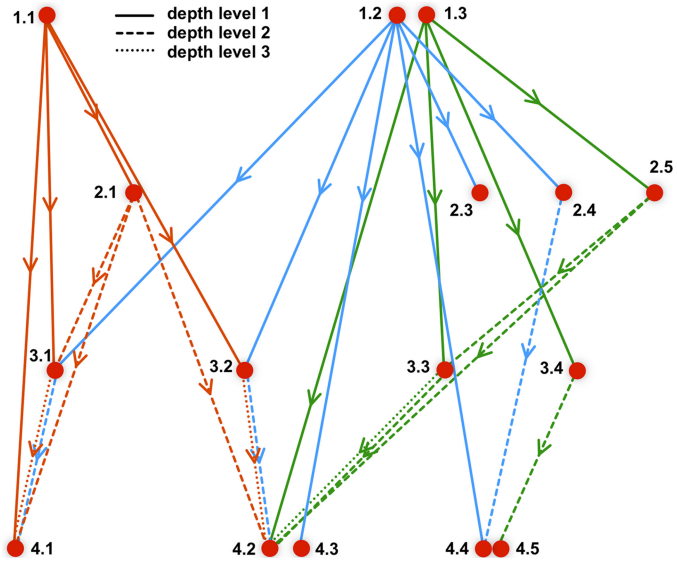

Fig. 8. Example of a depth-first path search used to retrieve non-singleton groups of vertices in a sparse SL compatibility graph. The example is using a sub-graph of the compatibility graph shown in Fig. 7, whose edges are made directive and hence the graph made directed. A separate search (denoted by a unique color) is started at each node in the highest layer, and multiple discoveries are allowed. By combining the retrieved paths along different depth levels according to the scheme outlined in Section III-B4, all relevant SL groups, potentially forming a clique in the undirected graph, are found.

Afterwards, each group is validated as a clique if any pair of vertices within it is connected (in the undirected graph, as opposed to the directed graph used for the path search), and conversely not validated if any such pair is not connected. Any non-validated groups are discarded, and within those remaining, any groups which are merely subsets of others are also discarded afterwards. The rationale behind this sifting is that when two perfectly-overlapping, *validated*, and differently-sized groups are available, it suffices to keep the bigger group. The subsets here can be discarded without losing genuine groups, because the validation of the clique property has already been performed. However, notwithstanding that the only remaining groups are cliques, these can still overlap at this stage.

5) SL Vertex Group Conflict Resolution: After the processing described in Section III-B4, further processing is required to ensure that the groups are distinct and do not overlap. Each SL vertex is forced to belong to only one group. This is because in the end, no SL should simultaneously belong to two TOA sets or sound wavefronts, provided SL detection with the linear Radon transform performs ideally — i.e., detects SLs corresponding to genuine sound wavefronts and detects those exclusively in a one-to-one manner.

First of all, the groups are sorted according to their 'tightness' — the average pairwise inter-circle distance between their vertices' SLs' corresponding circles. Each group is compared to every other group; when two groups share one or more vertices, these in addition to both groups are flagged as conflicting. The conflict resolution[11] then goes through the groups in order, from more to less tight, and inspects each of its conflicting groups in this same order: for every conflict, the conflicting vertices are assigned exclusively to the bigger group, i.e., kept in this bigger group and discarded from the smaller group, in the same rationale privileging bigger groups as in Section III-B4. This process is repeated sequentially until all conflicts are resolved. Finally, groups consisting of one single vertex are discarded.

[11]This scheme shares many principles with the neighborhood suppression scheme operating in the linear-Radon-transform-based SL detection in [55].

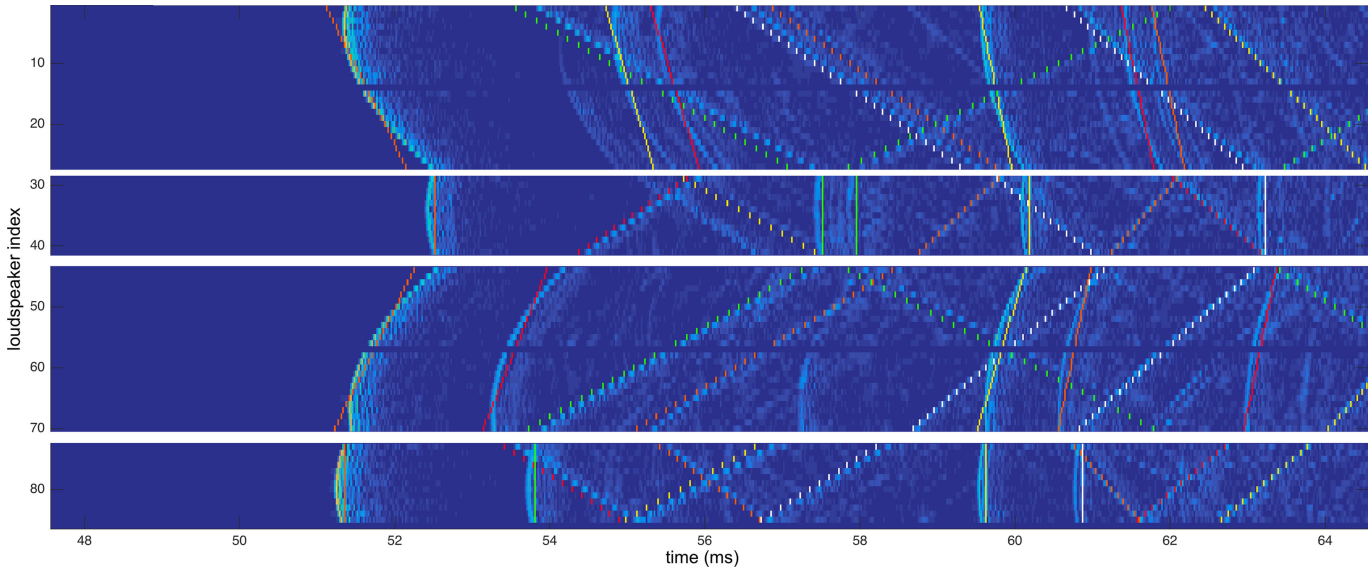

Fig. 9. Translation of detected SL groups into TOA sets (distinguished by color with repetition), onto the RIR stack in Fig. 2 (real measurements). Each SL in a group can be translated into a set of TOAs, which across sub-arrays form full 3D TOA sets.

C. Translation of Stack-Line Groups Into TOA Sets

The resulting SL groups are assembled into a set

$$\mathcal{G} = \Big\{ \big\{ (a,r)_g,\, r \in \{0..\widehat{R}-1\},\, a \in \{1..A\} \big\} : \forall g \Big\},$$

where (a,r) denotes the r-th SL on the a-th sub-array, and g indexes the groups. The corresponding TOA sets (Fig. 9) as formulated in Section II are disambiguated by construction according to

$$\big\{ \mathcal{T}_r \,:\, \forall r \in \{0\,..\,|\mathcal{G}|-1\} \big\} =$$

$$\Big\{ \bigcup_{a=1}^{A} \Big\{ \mathcal{T}_{ra} : \forall (a,r)_g \in \mathcal{G}(g) \Big\} : \forall g \in \{1..|\mathcal{G}|\} \Big\},$$

with $\mathcal{T}_r$ and $\mathcal{T}_{ra}$ as defined in (1) and (2), respectively; $|\mathcal{G}|$ denotes the total number[12] groups and $\cup$ denotes the union operation on sets.

IV. Real and Image Microphone Localization

After obtaining the disambiguated TOA sets $\{\mathcal{T}_r : \forall r \in \{0\,..\,|\mathcal{G}|-1\}\}$ according to the method in Section III, it becomes possible to perform geometrical interpretation of the RIRs. The first step in this regard is SSL.

The SSL scheme estimates the positions of the real and image microphones using spherical trilateration. Each TOA for a specific loudspeaker in a disambiguated set $\mathcal{T}_r$ is translated into a physical spatial range, which in turn defines a 3D sphere centered on the loudspeaker's position. When the loudspeaker array features enough geometrical diversity (Section I), these spheres intersect at a point[13] — instead of along a circle as in Fig. 5 — to determine the corresponding position to estimate.

For the r-th TOA set $\mathcal{T}_r$, the desired estimate $\widehat{\mathbf{i}}_r = (\widehat{i}_{r_x}, \widehat{i}_{r_y}, \widehat{i}_{r_z}, \widehat{d}^{\,2}_{\widehat{\mathbf{i}}_r})^T$ is obtained via $\widehat{\mathbf{i}}_r = \mathbf{A}^{+}\mathbf{b}$ [61], where $(.)^{+}$ denotes the matrix pseudo-inverse operator,

$$\mathbf{A} = \begin{bmatrix} -2s_{1x} & -2s_{1y} & -2s_{1z} & 1 \\ -2s_{2x} & -2s_{2y} & -2s_{2z} & 1 \\ \vdots & \vdots & \vdots & \vdots \\ -2s_{Lx} & -2s_{Ly} & -2s_{Lz} & 1 \end{bmatrix}, \mathbf{b} = \begin{bmatrix} d_{r1}^2 & -\|\mathbf{s}_1\|^2 \\ d_{r2}^2 & -\|\mathbf{s}_2\|^2 \\ \vdots & \vdots \\ d_{rL}^2 & -\|\mathbf{s}_L\|^2 \end{bmatrix},$$

$d_{rj} = c\,\tau_{rj}$ and $\widehat{d}^{\,2}_{\widehat{\mathbf{i}}_r}$ is an auxiliary variable corresponding to the squared norm estimate of $\widehat{\mathbf{i}}_r$. With the proposed approach, the ill-posedness is alleviated in both SSL and RL by assuming the position is always in the region in front of the array, i.e., enforcing a positive x coordinate $\widehat{i}_{r_x} = (\widehat{d}^{\,2}_{\widehat{\mathbf{i}}_r} - \widehat{i}^{\,2}_{r_y} - \widehat{i}^{\,2}_{r_z})^{1/2}$. The ill-posedness can occasionally cause the estimator to output $\widehat{d}^{\,2}_{\widehat{\mathbf{i}}_r}$, $\widehat{i}^{\,2}_{r_y}$ and $\widehat{i}^{\,2}_{r_z}$ such as the enforced x coordinate $\widehat{i}_{r_x}$ is complex, in which case the x coordinate of the estimated real microphone position[14] is used.

We further perform a few corrections made possible when the loudspeaker array is parallel to its nearest reflector[15] (Setups

[12]Note that R has been replaced by $|\mathcal{G}|-1$: this is because it is not guaranteed that the SL association perform flawlessly, not all the real or image reflectors' TOA sets might be detected, and it is possible that some incorrect TOA sets belonging to no such reflector are detected. More metrics on the performance of this detection are found in Section VI-B. Also note that while $\widehat{R}$ is the number of SLs in a single sub-array, the number of possible SL groups across sub-arrays is bigger (but still low thanks to sparsity): this number is in turn reduced to $\mathcal{G}$ after the processing in Section III-B.

[13]Since TOAs will include some noise, the intersection is not exact but approximate; hence, this intersection is computed in the least-squares sense, i.e., by minimizing the distances between the estimate and all the spheres.

[14]The real microphone position is easily distinguishable as the estimate obtained form the direct sound TOA set, i.e., the earliest TOA set, due to the direct line of sight.

[15]Deviations from this parallel placement are still allowed, but an error proportional to the discrepancy between the assumed and the real placement is to be expected. For the first of the corrections mentioned next, any deviation from the parallel placement results in an SSL error E_{ML} (Section VI-C) of $|2\sin(|\theta|)\widehat{i}_{r_x}|$, where θ is the array's vertical rotation angle around the z-axis. Therefore, this first correction is not used when the array is significantly rotated (as in Setup 6 in Section VI-A).

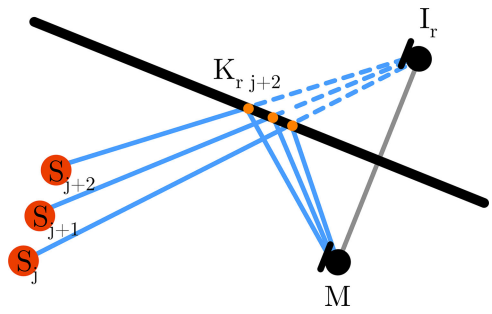

Fig. 10. Reflection point positions shown in detail for three consecutive loudspeakers j, $j+1$ and $j+2$. The thick black line represents the reflector under consideration, the blue lines represent the sound propagation paths, and the orange dots on the reflector represent reflection points. The letters S, M, K and I refer to a loudspeaker, real microphone, reflection point and image microphone, respectively.

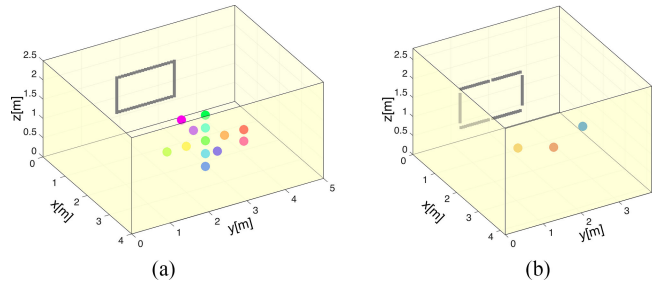

Fig. 11. Example microphone placement pattern for simulated and real measurements (Setups 1–6 and 7, respectively, in Tables I and II). The colored dots represent the microphone positions. The transparent beige surfaces represent the boundaries of the room, and the gray rectangle represents the loudspeaker array. (a) Simulated setup. (b) Real setup.

1–5 and 7 in Section VI-A). For instance, the image microphone nearest to the real microphone corresponds to the reflector nearest to the loudspeaker array. We invert the sign of the x coordinate for this image microphone exclusively, to obtain the physically-valid instead of the flipped position.

V. Reflector Localization

An interim step between SSL and RL is the computation of sets of reflection points for each reflector. Each of the multiple image microphone positions ($\mathbf{i}_r : \forall r$), obtained by means of SSL (Section IV), corresponds to a specific reflector and its L (real microphone, image microphone, loudspeaker) triplets. This shows that the RL problem involving multiple reflectors consists of several *sub*-problems involving only a single reflector. For each image microphone, the method outlined in [22] is applicable[16] to obtain a set of reflection points $\{\mathrm{K}_{rj} : \forall j \in \{1..L\}\}$ (Fig. 10). The reflector is the plane passing through all the reflection points; hence, all these points satisfy the reflector's equation, and the reflector's parameters are determined with

$$\mathbf{n}_r = \underset{\mathbf{n}_r}{\operatorname{argmin}} \sum_{j=1}^{L} \langle \mathbf{n}_r, \mathbf{k}_{rj} \rangle^2 \,,\; o_r = -\langle \mathbf{n}_r, \mathbf{k}_{r1} \rangle \,,$$

where $\mathbf{k}_{rj}$ is the position of the reflection point K_{rj}.

VI. Performance Evaluation

This section presents a twofold evaluation of IMRL: the 3D TOA detector and disambiguator (Section III) and SSL-RL (Sections IV and V) are tested separately. Both evaluations operated on the same data sets, which contained simulated as well as measured RIRs. For consistency and comparison purposes, the same evaluation frameworks, metrics and setup configurations as in the present authors' previous contributions [22], [55] were used.

[16] With a small adaptation: extending the coordinate system $\mathcal{C}_{\mathrm{S}j}$ [22], centered on the j-th loudspeaker, to 3D.

A. Setups and Data Sets

Each setup (one room and one loudspeaker array arrangement) with a specific microphone position formed an independent setting with a multi-sub-array RIR stack on which IMRL was applied. The dimensions of the rooms are shown in Table II. The presence of multiple microphone positions enabled more data points, as for each setup the metrics were averaged across all such settings in Sections VI-B and VI-C, in addition to being averaged across reflectors or TOA sets. The sampling frequency used in all setting was 48 kHz.

1) Simulated RIRs: Six different setups were prepared with RIRs simulated using [68] with the image method [42] of seventh order. A setup consisted of a shoebox room of specific dimensions and one 1.5 × 0.9 m uniform rectangular loudspeaker array with 64 drivers, divided into 4 linear sub-arrays. It included 1 arbitrarily-positioned omnidirectional microphone and 12 other omnidirectional microphones which were placed on a 3D-cross pattern centered around the farther half of each room (Fig. 11(a)) 1–6 m away from the loudspeaker array, which was placed 0.3 m away from the nearest wall. In Setup 4 the pattern was centered in the middle of the room instead. The attenuation coefficients α_{rj} were set to 0.5 for all reflectors. While Setups 1–5 had the array parallel to one of the reflectors, Setup 6 had the array rotated 15° counter-clockwise w.r.t its center around the z-axis, to test the generalizability of IMRL.

In addition to the 6 exclusively-simulated setups, a simulated replica of a seventh setup (involving real measurements) was prepared. This setup was intended as a means to directly compare the performance between real and simulated RIRs under the same geometrical conditions. With this additional setup, a total number of 81 simulated settings were evaluated.

2) Measured RIRs: RIRs were measured using swept-sine emission [69] for each (loudspeaker, microphone) pair in a setup involving a 1.56 × 0.96 m uniform rectangular loudspeaker array, divided into 4 linear sub-arrays. We used multiple uniform linear arrays of 13 drivers each to perform measurements at 78 loudspeaker positions; some other positions were not measured because, for practical reasons, these uniform linear arrays cannot reconstruct a uniform rectangular array perfectly (see white overlay on Fig. 2). This setup included 3 omnidirectional microphones, forming 3 settings, placed across the farther half of a shoebox room (Fig. 11(b)) 2.5–3.5 m away from the loud-

speaker array, which was placed 0.5 m away from the nearest wall. The hardware latency was provided to the algorithm based on a comparison with the aforementioned simulated replica. Average reverberation time (RT60) was estimated at 0.45 s. This setup provided the data for Figs. 2 and 4–9.

B. TOA Detector and Disambiguator Evaluation

The performance of the TOA detector and disambiguator (Fig. 9, Section III) was objectively assessed with three metrics: the true positive rate indicating the percentage of detected TOA sets that match reference TOA sets, the false discovery rate indicating the percentage of detected TOA sets that do not match reference TOA sets and the error of the correctly detected TOA sets given by the root mean square error (RMSE) between a set's TOAs and their matched reference TOAs. Each detected TOA set was compared to all reference TOA sets, and counted as correct when a one-to-one match with an RMSE of 0.5 ms or less was found. The reference TOAs were retrieved from 3D simulations using the seventh-order image model [42]. Higher-order TOAs, and those beyond the truncation time, were not considered in the evaluation. Better performance is indicated by higher true positive rates, lower false discovery rates and lower RMSEs.

The same SL detection parameters as in [55] were used, with the exception of the new relative threshold $T_r = 15\%$ and the maximum number of SLs which was increased[17] to 20. As for SL association parameters, the tolerances $T_c = 0.02$, $T_{\widetilde{d}_{icc}} = 5$ cm (Annex A) and $T_{d_{ic}} = 30$ cm were used.

C. SSL-RL Evaluation

To assess the accuracy of the real and image microphone position (SSL) estimates, the Euclidean distance $E_{ML} = \|\mathbf{m} - \widehat{\mathbf{m}}\|$ [10] between the true ($\mathbf{m}$) and estimated ($\widehat{\mathbf{m}}$) microphone positions was used. To assess the accuracy of reflection point localization (and RL), the orientation error $O_{RL} = |\arccos(\langle \mathbf{n}_r, \widehat{\mathbf{n}}_r \rangle)|$ [12] between the true ($\mathbf{n}_r$) and estimated ($\widehat{\mathbf{n}}_r$) reflectors' normal vectors was used; additionally, the offset $D_{RL} = \big|\,|\langle \mathbf{n}_r, (\mathbf{m} - \mathbf{x}) \rangle| - |\langle \widehat{\mathbf{n}}_r, (\mathbf{m} - \widehat{\mathbf{x}}) \rangle|\,\big|$ [12] in terms of the distance of the true and estimated reflectors to the true real microphone location $\mathbf{m}$ was used, where $\mathbf{x}$ and $\widehat{\mathbf{x}}$ represent points on the true and estimated reflectors, respectively. Only first-order image microphones and physical reflectors were considered; lower metrics (Fig. 12) indicate better performance. We also computed the average number of estimated physical reflectors, and the percentage of settings where this number reached 6 (corresponding to 'full' RGI).

D. Results and Discussion

The TOA detector and disambiguator metrics, shown in Table I, were assembled separately for simulated reference TOA sets of orders 0–2, to facilitate the correct interpretation of the results. The direct-sound TOA set was always correctly detected (zero-order true positive rates of 100%), regardless of the setup or setting. Additionally, the method achieved nearly-perfect first-order true positive rates (92–100%, averaging 98.2%) on the simulated setups (rows '1–7 (S)' in Table I); as to the real measurement setup (row '7 (R)' in Table I), one of its six reflectors produced quasi-invisible SLs due to loudspeaker directivity; thus, the 83.3% first-order true positive rate for first-order SLs in this setup can also be effectively considered a 100% true positive rate. The true positive rates decrease with increasing reflection order, considering that reflections become denser for higher orders and later TOAs, rendering correct TOA detection and disambiguation more challenging. However, this should not be viewed as a limitation, as only first-order reflections are critical for RGI applications. Interestingly, the proposed method can still detect a significant amount of second- and higher- order reflections. Moreover, the false discovery rates were reasonable in both simulated and real setups (4–15%, averaging 8.8% and 12.7% across simulated and real setups, respectively) and RMSE values were similarly low (80–160 μs, averaging 104.5 μs and 168.4 μs for first-order TOA sets on simulated and real setups, respectively). The accuracy of the method and its robustness to noise can be observed in Fig. 9.

For real data, the comparison between detected and reference TOA sets is inherently challenging, because any small discrepancy between the modeled, measured and actual loudspeaker positions — and consequently the corresponding TOAs — doubles with the reflection order. This makes the matching between detected and reference TOA sets difficult for higher orders, and

[17]Both of these changes are a means to account for the increased reflection density, which is due to the extension from 2D to 3D.

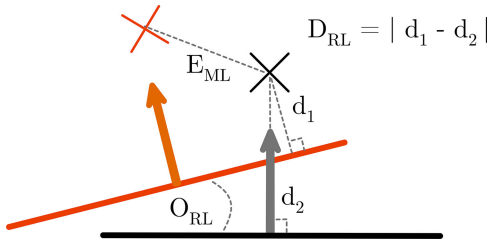

Fig. 12. Visual representation of RL error metrics. The black line, arrow and cross indicate a true reflector, its normal vector and image microphone position, while the red line, arrow and cross indicate their estimated counterparts.

TABLE I
OBTAINED TOA DETECTION AND DISAMBIGUATION PERFORMANCE METRICS (SECTION VI-B)

Order	All	0 (direct sound)		1		2	
Setup	FDR%	TPR%	RMSE μs	TPR%	RMSE μs	TPR%	RMSE μs
1 (S)	10.3	100	95.8	92.3	121.8	61.5	132.6
2 (S)	6.0	100	82.00	97.4	91.6	63.7	100.9
3 (S)	7.7	100	81.0	100	95.2	61.1	94.6
4 (S)	4.3	100	81.0	100	88.2	65.6	93.4
5 (S)	7.6	100	84.4	100	101.2	56.3	101.8
6 (S)	14.7	100	96.0	97.4	113.3	58.1	109.5
7 (S)	10.8	100	108.3	100	120.1	63.0	156.2
Avg (S)	**8.8**	**100**	**89.8**	**98.2**	**104.5**	**61.3**	**112.7**
7 (R)	**12.7**	**100**	**142.1**	**83.3**	**168.4**	**33.3**	**154.2**

'(S)' indicates simulations, '(R)' real measurements, and 'Avg (S)' the simulations' average metrics; FDR denotes the false discovery rate, TPR indicates the true positive rate.

TABLE II
SSL-RL PERFORMANCE METRICS FOR 7 SIMULATED SETUPS (ROWS 1–7) AND 1 REAL SETUP (ROW 8)

Setup	Room dimensions (m)	$E_{ML}(cm)$	$D_{RL}(cm)$	$O_{RL}(degrees)$	% of full RGI	average number of est. refl.
1	$4.5 \times 5 \times 2.5$	20.494 ± 22.530	7.398 ± 6.794	2.974 ± 2.557	84.616	5.846
2	$6 \times 4 \times 2.9$	17.642 ± 13.837	7.615 ± 6.531	2.715 ± 2.179	84.616	5.846
3	$6 \times 8.5 \times 3.5$	24.206 ± 26.591	8.201 ± 6.833	2.408 ± 1.995	92.301	5.923
4	$9 \times 7.5 \times 3$	22.020 ± 23.120	6.377 ± 5.379	2.228 ± 1.882	100	6
5	$6 \times 12 \times 3$	29.387 ± 33.507	8.187 ± 7.293	2.308 ± 1.861	100	6
6	$4.5 \times 5 \times 3$, array rotated by 15°	7.936 ± 10.563	4.219 ± 4.932	1.276 ± 1.151	N/A	4.616
7 (S)	$4.26 \times 3.87 \times 2.80$, simulated RIRs	14.338 ± 13.524	7.248 ± 7.304	2.290 ± 1.619	100	6
7 (R)	$4.26 \times 3.87 \times 2.80$, measured RIRs	20.973 ± 21.628	9.055 ± 6.784	3.545 ± 2.974	N/A	4.667

Shown are the metrics (Section VI-C) averaged across all microphone positions and reflectors, with a mean value followed by $\pm$ the standard deviation.

has more problematic implications in 3D than in 2D. More precisely, true positive rate values for orders 0–2 are negligibly affected; however, the false discovery rate is computed across reflection orders 0–7 and should be interpreted accordingly: it would be lower if all detected TOA sets could be correctly matched with reference TOA sets.

The results of the SSL-RL evaluation are shown in Table II. The SSL errors E_{ML} of 8–30 cm and 21 cm, for simulated and real setups respectively, were reasonable; however they translate into smaller RL errors D_{RL} of 4–8 and 9 cm, and O_{RL} errors of 1–3° and 3.5°; for simulated and real setups, respectively. This is predictable because, depending on their orientations, SSL errors can cancel each other out in terms of final RL errors: for example, for a real and an image microphone on either side of a reflector, equal SSL errors going in opposite orientations along the normal of that reflector will not result in any RL error. In general, the proposed approach performs similarly to what is reported in the literature [13], [18], [32], [40], [44], [51], in spite of the ill-posedness. The metrics are inferior to those in [22] because [22] allowed supervised TOA disambiguation, whereas in this paper we used automated TOA disambiguation. An example of final RGI results with simulated RIRs is displayed in Fig. 13, for Setup 7 (S) (the $4.26 \times 3.87 \times 2.80$ m room). In this specific case, TOA detection and disambiguation achieved true positive rates of 100% and 100%; and RMSEs of 110.5 μs and 119.1 μs; for direct and first-order TOA sets, respectively; in addition, a low false discovery rate of 9.52% was achieved. In other words, all TOA sets corresponding to physical reflectors in the room were detected with low RMSEs. The resulting SSL-RL errors were $E_{ML} = 17.73$ cm, $D_{RL} = 9.24$ cm and $O_{RL} = 2.61°$.

Our TOA processing features a TOA estimation approach based on SL detection. The approximation of wavefronts with SLs is largely sufficient for disambiguation purposes, but it lacks accuracy for high-precision RL. The early wavefronts in an RIR stack present noticeably more curvature than the later wavefronts (Fig. 2), due to the smaller distance between their corresponding real or image microphones and the loudspeaker array. After associating SLs across sub-arrays, the mismatch between the wavefronts' slight curvature and their corresponding detected SLs' linearity can be problematic. This is because for the final TOA sets, large TOA discontinuities appear at the borders between sub-arrays (see earliest TOA set in Fig. 9). Nonetheless, this is not a significant problem considering that TOA detection in individual RIRs can be used interim and its results merged with the associated SLs based on a nearest-neighbor logic. This

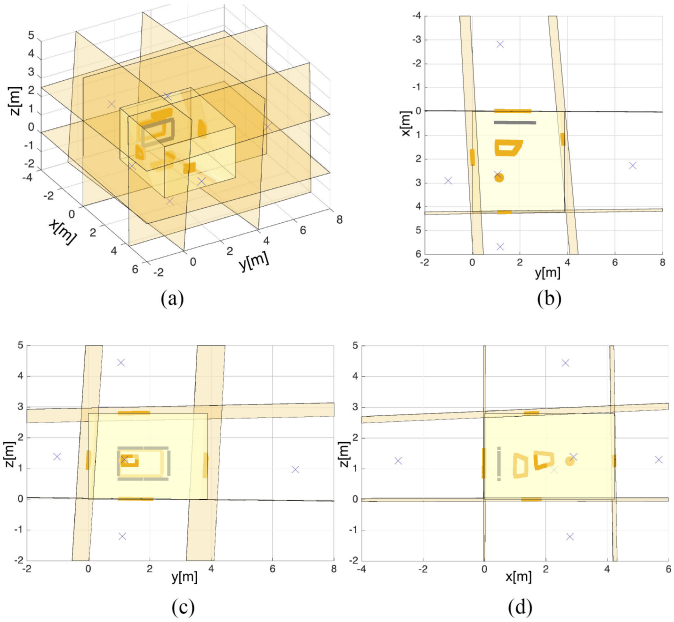

Fig. 13. Example RL estimation result for simulated RIRs (Setup 7 (S) in Table II, third microphone position). The transparent beige surfaces represent the boundaries of the room, and the gray rectangle represents the loudspeaker array; yellow projections of the array represent the computed reflection points on the reflectors, which are in turn represented by the transparent yellow planes. The real microphone position is represented by a yellow dot, and the image microphone positions by blue crosses. Obstructing reflectors were omitted in the flattened 2D views, namely: the floor and ceiling reflectors in the top view (13(b)), the front and back reflectors in the frontal view (13(c)) and finally the left and right reflectors in the side view (13(d)). This is to facilitate the visual inspection of the results. (a) 3D view. (b) Top view. (c) Frontal view. (d) Side (left) view.

or other approaches can be adopted to improve accuracy, for example by increasing the operating sampling frequency of the whole system. We chose to accept a loss in accuracy in order to give a proof-of-concept that RGI is possible without TOA detection in individual RIRs.

One considerable advantage with our approach is its feasibility on often imperfect, noisy and generally challenging measured RIRs. In fact, the performance losses are moderate between simulated and real setup equivalents (compare rows '7 (S)' and '7 (R)' in Tables I and II). We also consider the minimal use of optimization in our approach as an advantage — the only optimization is performed in SSL, and only in its efficient least-squares variant. The essence of the algorithm is analytical, in addition to being parallelizable.

While the algorithm could not localize all reflectors in some settings, this was achieved in the majority of settings. When it was not possible, the limitation was often due to extraneous circumstances: namely a missing array dimension in the loudspeaker array and nearly-absent front wall reflections in the case of real measurements, due to loudspeaker directivity. The measures we adopted to counter the missing array dimension work under the condition that the loudspeaker array be placed near and approximately-parallel to a reflector. Nonetheless, in common measurement setting, no reason should hinder such a placement.

More generally, we expect the algorithm to work independently of whether the room is shoebox-shaped or not. One prevalent and unavoidable assumption in all RGI algorithms, however, is that reflections from all reflectors need to be audible to the array. In other words, if TOA peaks do not exist in the measured RIRs, it is not possible to detect them and therefore the problem is ill-defined. Provided reflections from all reflectors in such non-shoebox room geometries are audible to the array, the only impact such geometries would have lies in the front/back ambiguity. If the geometry of such a room places more first order image microphones behind the array than only the image microphone of the wall near the array, then these will be estimated with an incorrect, flipped x coordinate; this is nonetheless due to our limited array geometry.

Finally, the synchronization requirement for IMRL can be alleviated by running dedicated spatio-temporal synchronization before SL detection. If needed, this can be achieved by running a TDOA-based SSL scheme such as [56] to estimate the position of the real microphone. This entails distinguishing and estimating the direct sound TDOAs, but this task is feasible due to the salience, curvature and inherent distinguishability of the direct sound wavefront. Nonetheless, a high sampling frequency is required to avoid added synchronization errors, as these can lead to a larger perspective error in room geometry reconstruction.

VII. Conclusion

We presented a largely automated RGI method to estimate the positions and orientations of reflectors in shoebox rooms, without it being limited to such rooms. In contrast to the state-of-the-art, the presented method relies on image processing, by approximating wavefronts with lines on images, coupled with graph-theory for the core task of TOA detection and disambiguation. The method formulates most of the remaining tasks analytically. It uses relatively nonrestrictive transducer placement and synchronization assumptions to counteract the ill-posedness of a limited transducer setup, while achieving good accuracy and remaining computationally feasible, especially when using measured RIRs. The room geometry reconstruction after reflector localization is the only step that was performed in a supervised manner.

Appendix A
Efficient and Robust Computation of the Distance Between Circles in 3D

In the following, for brevity, the indexing $(.)_{1/2}$ designates two instances of an entity. Consider two circles $\mathrm{C}_{1/2}$ and their encompassing 3D planes $\mathrm{P}_{1/2}$, they are characterized by

$$\mathrm{C}_{1/2} = \begin{cases} \mathbf{c}_{1/2} \\ \mathbf{n}_{a,1/2} \\ d_{ca,1/2} \end{cases}, \mathrm{P}_{1/2} = \begin{cases} \mathbf{c}_{1/2} \\ \mathbf{v}_{1,1/2} \\ \mathbf{v}_{2,1/2} = \mathbf{v}_{1,1/2} \times \mathbf{n}_{a,1/2} \end{cases},$$

where $\mathbf{c}_{1/2}$, $\mathbf{n}_{a,1/2}$ and $d_{ca,1/2}$ characterize the circles' centers, normal vectors and radii, respectively; $\times$ denotes the cross product between vectors and $\mathbf{v}_{1,1/2}$ is any arbitrary unit vector orthogonal to $\mathbf{n}_{a,1/2}$; the direction vectors $\mathbf{v}_{1,1/2}$ and $\mathbf{v}_{2,1/2}$ form orthonormal bases for their planes. An example of a circle and its corresponding plane, in addition to the relevant vectors and quantities, is shown in Fig. 14.

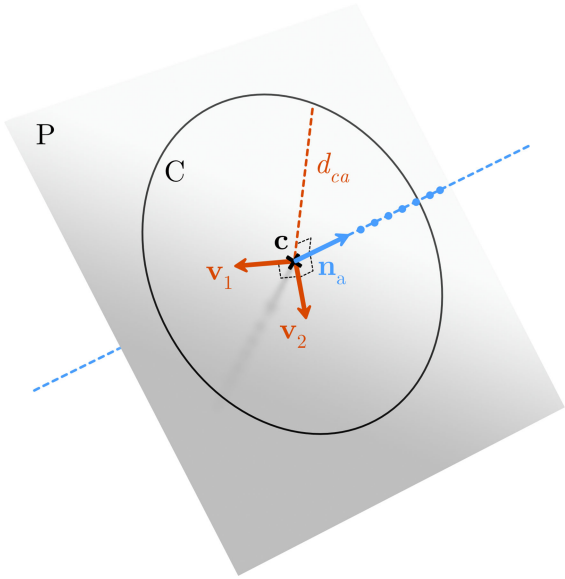

Fig. 14. 3D view of a circle, its corresponding plane and the relevant vectors and quantities. The plane P (gray plane) encompasses the circle C (black circle), which has radius d_{ca} and center $\mathbf{c}$. Orange indicates objects residing inside P, while blue indicates objects residing outside P. In our application, a linear subarray (denoted by a series of blue dots) determines the normal vector $\mathbf{n}_a$ of the plane and circle. The vectors $\mathbf{v}_1$ and $\mathbf{v}_2$, forming a basis for P, are orthonormal, and are both orthogonal to $\mathbf{n}_a$.

The procedure for computing the distance between the circles is carried out differently according to a hierarchy of cases. Because the circles' orientations are determined by their respective normal vectors, the circles' planes either intersect or not depending on whether the vectors are collinear or non-collinear. The differentiation is based on the observation that perfectly collinear unit vectors have a scalar product equal to 1 or -1. This can be distinguished via

$$\left| \left| \left\langle \mathbf{n}_{a,1}, \mathbf{n}_{a,2} \right\rangle \right| - 1 \right| \leq T_c, \tag{4}$$

where T_c denotes a collinearity tolerance close to zero.[18]

A. Collinear Case

If the normal axes are collinear according to (4), the planes P_1 and P_2 are parallel. The inter-circle distance d_{ic} is computed in a

[18]The use of tolerances is meant to account for possible errors in estimated vectors and quantities.

two-step procedure: first, the circles are projected to a common 2D plane (P_1, using its orthonormal basis $\{\mathbf{v}_{1,1}, \mathbf{v}_{2,1}\}$), the 2D inter-circle distance $\widetilde{d}_{ic}$ between their projections is computed within that plane; second, the orthogonal inter-plane distance d_{ip} between P_1 and P_2 is taken into account. The application of this projection on an entity is indicated by a $\widetilde{\cdot}$. The 2D projections of $\mathrm{C}_{1/2}$ onto P_1 are respectively

$$\widetilde{\mathrm{C}}_1 = \begin{cases} \widetilde{\mathbf{c}}_1 = (0,0)^T \\ d_{ca,1} \end{cases}, \widetilde{\mathrm{C}}_2 = \begin{cases} \widetilde{\mathbf{c}}_2 = \begin{bmatrix} \mathbf{v}_{1,1}{}^T \\ \mathbf{v}_{2,1}{}^T \end{bmatrix} \cdot \mathbf{c}_2 \\ d_{ca,2} \end{cases}, \quad (5)$$

where $\widetilde{\mathbf{c}}_{1/2}$ is the orthonormal projection of $\mathbf{c}_{1/2}$ onto P_1. Because the planes P_1 and P_2 are parallel when the relevant normal vectors are collinear, both radii are conserved, and the 3D circle C_2 retains its circular shape when projected, instead of degenerating into an ellipse.

After this projection, the 2D inter-circle distance $\widetilde{d}_{ic}$ is computed using the 2D inter-circle-center distance $\widetilde{d}_{icc} = \|\widetilde{\mathbf{c}}_1 - \widetilde{\mathbf{c}}_2\|$ between the 2D circles' centers, and the radii $d_{ca,1}$ and $d_{ca,2}$, according to the following sub-cases [66, Section 5.1]:

- If $\widetilde{d}_{icc} \leq T_{\widetilde{d}_{icc}}$: the 2D circles are concentric if their centers lie within an inter-circle-center distance tolerance $T_{\widetilde{d}_{icc}}$, thus; $\widetilde{d}_{ic} = |d_{ca,1} - d_{ca,2}|$.
- Otherwise if the 2D circles cannot be considered concentric:
  - If $\widetilde{d}_{icc} \geq d_{ca,1} + d_{ca,2}$, the 2D circles lie completely apart at a distance $\widetilde{d}_{ic} = \widetilde{d}_{icc} - (d_{ca,1} + d_{ca,2})$.
  - Otherwise :
    - If $d_{ca,1} \neq d_{ca,2}$ (the radii are unequal) :
      - If $(\widetilde{d}_{icc} + d_{ca,2} \geq d_{ca,1}) \wedge (\widetilde{d}_{icc} - d_{ca,2} \geq -d_{ca,1})$, the smaller 2D circle is not fully contained inside the bigger 2D circle, but lies partially inside and partially outside, therefore; the circles intersect and $\widetilde{d}_{ic} = 0$.
      - Otherwise one 2D circle is contained in the other and $\widetilde{d}_{ic} = |d_{ca,1} - d_{ca,2}| - \widetilde{d}_{icc}$.
    - Otherwise the radii are equal, $\widetilde{d}_{icc} + d_{ca,2} \geq d_{ca,1} \iff \widetilde{d}_{icc} \geq 0$, and since $\widetilde{d}_{icc}$ is always a positive distance, the circles intersect (as above) and $\widetilde{d}_{ic} = 0$.

Considering that the planes are parallel, the inter-plane distance d_{ip} is computed by choosing any point in P_2, projecting it orthogonally onto P_1, similarly to $\widetilde{\mathbf{c}}_2$ in (5), and then computing the distance between the chosen point and its projection. The distances d_{ip} and $\widetilde{d}_{ic}$ then determine the distance d_{ic} between the 3D circles C_1 and C_2 via Pythagoras' theorem with $d_{ic} = (\widetilde{d}_{ic}^2 + d_{ip}^2)^{1/2}$.

B. Non-Collinear Case

If the normal vectors are non-collinear according to (4), the planes P_1 and P_2 cannot be parallel and therefore intersect along an infinite line. Accordingly, this line is computed and in turn intersected separately with the circles C_1 and C_2. The distance between these separate individual intersections is d_{ic}.

The intersection line is parametrized by a direction vector $\mathbf{n}_{a,1} \times \mathbf{n}_{a,2}$ and a characteristic point [70]:

$$\mathbf{l}_0 = \Big\{ \big(\langle \mathbf{n}_{a,1}, \mathbf{c}_1 \rangle \|\mathbf{n}_{a,2}\|^2 - \langle \mathbf{n}_{a,2}, \mathbf{c}_2 \rangle \langle \mathbf{n}_{a,1}, \mathbf{n}_{a,2} \rangle\big) \mathbf{n}_{a,1} + \big(\langle \mathbf{n}_{a,2}, \mathbf{c}_2 \rangle \|\mathbf{n}_{a,1}\|^2 - \langle \mathbf{n}_{a,1}, \mathbf{c}_1 \rangle \langle \mathbf{n}_{a,1}, \mathbf{n}_{a,2} \rangle\big) \mathbf{n}_{a,2} \Big\} \Big/ \Big\{ \|\mathbf{n}_{a,1}\|^2 \|\mathbf{n}_{a,2}\|^2 - \langle \mathbf{n}_{a,1}, \mathbf{n}_{a,2} \rangle^2 \Big\}.$$

Intersecting this line with the i-th circle C_i ($i \in \{1,2\}$) is equivalent to intersecting it with the sphere centered on $\mathbf{c}_i$ with radius $d_{ca,i}$, which is done by solving the quadratic equation [70]

$$\|\mathbf{n}_{a,1} \times \mathbf{n}_{a,2}\|^2 u_i{}^2 + 2\big\langle (\mathbf{l}_0 - \mathbf{c}_i), (\mathbf{n}_{a,1} \times \mathbf{n}_{a,2}) \big\rangle u_i + \|\mathbf{l}_0 - \mathbf{c}_i\|^2 - d_{ca,i}^2 = 0$$

for u_i. The intersection is given by $\mathbf{m}_i = \mathbf{l}_0 + u_i(\mathbf{n}_{a,1} \times \mathbf{n}_{a,2})$. As a convention suited to our application, the solutions yielding a positive x coordinate are chosen, in light of the presence of two solutions for each intersection, but this is irrelevant for the purpose of measuring the distance d_{ic} when applied consistently. After intersecting the line with each sphere, the inter-circle distance is given by $d_{ic} = \|\mathbf{m}_1 - \mathbf{m}_2\|$. For the general case, it suffices to compute the minimal distance across both pairs of intersection points from each of the circles.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their constructive feedback on the paper. Acknowledgement is also due to the authors' colleagues for their help in the editing process.

REFERENCES

[1] D. de Vries and M. Boone, "Wave field synthesis and analysis using array technology," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 1999, pp. 15–18.
[2] A. Canclini, D. Marković, F. Antonacci, A. Sarti, and S. Tubaro, "A room-compensated virtual surround system exploiting early reflections in a reverberant room," in *Proc. Eur. Signal Process. Conf.*, Aug. 2012, pp. 1029–1033.
[3] A. Ando and M. Fujii, "Control of frame loudspeaker array for 3-D television," in *Proc. Audio Eng. Soc. Conv*. New York, NY, USA: Audio Engineering Society, Apr. 2014.
[4] P. Annibale *et al.*, "The SCENIC project: Space-time audio processing for environment-aware acoustic sensing and rendering," in *Proc. Audio Eng. Soc. Conv.*, Oct. 2011, pp. 1–10.
[5] F. Ribeiro, D. Ba, C. Zhang, and D. Florêncio, "Turning enemies into friends: Using reflections to improve source source localization," in *Proc. Int. Conf. Multimedia Expo.*, Singapore, Jul. 2010, pp. 731–736.
[6] A. Plinge and G. A. Fink, "Multi-speaker tracking using multiple distributed microphone arrays," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Florence, Italy, May 2014, pp. 614–618.
[7] Y. Peled and B. Rafaely, "Method for dereverberation and noise reduction using spherical microphone arrays," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2010, pp. 113–116.
[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.*, Mar. 2010, pp. 2822–2825.
[9] 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.

[10] J. Filos, E. A. P. Habets, and P. A. Naylor, "A two-step approach to blindly infer room geometries," in *Proc. Int. Workshop Acoust. Echo Noise Control*, Tel Aviv, Israel, Sep. 2010, pp. 1–4.
[11] J. Filos, A. Canclini, M. R. P. Thomas, F. Antonacci, A. Sarti, and P. A. Naylor, "Robust inference of room getometry from acoustic impulse responses," in *Proc. Eur. Signal Process. Conf.*, Barcelona, Spain, Aug. 2011, pp. 161–165.
[12] J. Filos, A. Canclini, F. Antonacci, A. Sarti, and P. Naylor, "Localization of planar acoustic reflectors from the combination of linear estimates," in *Proc. Eur. Signal Process. Conf.*, Aug. 2012, pp. 1019–1023.
[13] E. Nastasia, F. Antonacci, A. Sarti, and S. Tubaro, "Localization of planar acoustic reflectors through emission of controlled stimuli," in *Proc. Eur. Signal Process. Conf.*, 2011, pp. 156–160.
[14] P. Annibale and R. Rabenstein, "Accuracy of time-difference-of-arrival based source localization algorithms under temperature variations," in *Proc. Int. Symp. Control, Commmun. Signal Process.*, Mar. 2010, pp. 1–4.
[15] P. Annibale, J. Filos, P. Naylor, and R. Rabenstein, "Geometric inference of the room geometry under temperature variations," in *Proc. Int. Symp. Control, Commmun. Signal Process.*, May 2012, pp. 1–4.
[16] A. Canclini, P. Annibale, F. Antonacci, A. Sarti, R. Rabenstein, and S. Tubaro, "From direction of arrival estimates to localization of planar reflectors in a two dimensional geometry," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2011, pp. 2620–2623.
[17] A. Canclini *et al.*, "Exact localization of acoustic reflectors from quadratic constraints," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoustics*, 2011, pp. 17–20.
[18] L. Remaggi, P. J. B. Jackson, P. Coleman, and W. Wang, "Room boundary estimation from acoustic room impulse responses," in *Proc. Sensor Signal Process. Defence*, Sep. 2014, pp. 1–5.
[19] 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.
[20] L. Remaggi, P. J. B. Jackson, and P. Coleman, "Source, sensor and reflector position estimation from acoustical room impulse responses," in *Proc. 22nd Int. Congr. Sound Vib.*, 2015, pp. 1–8.
[21] 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.
[22] Y. El Baba, A. Walther, and E. A. P. Habets, "Reflector localization based on multiple reflection points," in *Proc. Eur. Signal Process. Conf.*, Budapest, Hungary, Aug. 2016, pp. 1458–1462.
[23] H. Naseri and V. Koivunen, "Cooperative simultaneous localization and mapping by exploiting multipath propagation," *IEEE Trans. Signal Process.*, vol. 65, no. 1, pp. 200–211, Jan. 2017.
[24] R. Schmidt, "A new approach to geometry of range difference location," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 8, no. 6, pp. 821–835, Nov. 1972.
[25] 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.
[26] A. Moore, M. Brookes, and P. Naylor, "Room geometry estimation from a single channel acoustic impulse response," in *Proc. Eur. Signal Process. Conf.*, Sep. 2013, pp. 1–5.
[27] A. O'Donovan, R. Duraiswami, and D. Zotkin, "Imaging concert hall acoustics using visual and audio cameras," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2008, pp. 5284–5287.
[28] 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.
[29] S. Tervo, J. Pätynen, and T. Lokki, "Acoustic reflection path tracing using a highly directional loudspeaker," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2009, pp. 245–248.
[30] E. V. Lancker, "Localization of reflections in auditoriums using time delay estimation," in *Proc. Audio Eng. Soc. Conv.*, Feb. 2000, pp. 1–8.
[31] S. Tervo and T. Korhonen, "Estimation of reflective surfaces from continuous signals," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2010, pp. 153–156.
[32] S. Tervo and T. Tossavainen, "3-D room geometry estimation from room impulse responses," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2012, pp. 513–516.
[33] D. Ba, F. Ribeiro, C. Zhang, and D. Florêncio, "L1 regularized room modeling with compact microphone arrays," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2010, pp. 157–160.
[34] F. Ribeiro, C. Zhang, D. A. Florencio, and D. 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.
[35] D. Aprea, F. Antonacci, A. Sarti, and S. Tubaro, "Acoustic reconstruction of the geometry of an environment through acquisition of a controlled emission," in *Proc. 17th Eur. Signal Process. Conf.*, 2009, pp. 710–714.
[36] I. Dokmanic, R. Parhizkar, A. Walther, Y. M. Lu, and M. Vetterli, "Acoustic echoes reveal room shape," *Proc. Nat. Acad. Sci.*, vol. 110, no. 30, pp. 12186–12191, 2013.
[37] R. S. Bandaru, A. Sørnes, J. D'hooge, and E. Samset, "2D localization of specular reflections using ultrasound," in *Proc. 2014 IEEE Int. Ultrason. Symp.*, Sep. 2014, pp. 2209–2212.
[38] L. Zamaninezhad, P. Annibale, and R. Rabenstein, "Localization of environmental reflectors from a single measured transfer function," in *Proc. Int. Symp. Control, Commmun. Signal Process.*, May 2014, pp. 157–160.
[39] M. Krekovic, I. Dokmanic, and M. Vetterli, "EchoSLAM: Simultaneous localization and mapping with acoustic echoes," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2016, pp. 11–15.
[40] I. Dokmanic, L. Daudet, and M. Vetterli, "From acoustic room reconstruction to SLAM," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2016, pp. 6345–6349.
[41] T. Nowakowski, N. Bertin, R. Gribonval, J. de Rosny, and L. Daudet, "Membrane shape and boundary conditions estimation using eigenmode decomposition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2016, pp. 3336–3340.
[42] 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, Apr. 1979.
[43] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.
[44] M. Krekovic, I. Dokmanic, and M. Vetterli, "Omnidirectional bats, point-to-plane distances and the price of uniqueness," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2017, pp. 3261–3265.
[45] M. Kuster and D. de Vries, "Modelling and order of acoustic transfer functions due to reflections from augmented objects," *EURASIP J. Appl. Signal Process.*, vol. 2007, no. 1, pp. 182–182, Jan. 2007.
[46] I. Dokmanic, Y. M. Lu, and M. Vetterli, "Can one hear the shape of a room: The 2-D polygonal case," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2011, pp. 321–324.
[47] D. Markovic, F. Antonacci, A. Sarti, and S. Tubaro, "Estimation of room dimensions from a single impulse response," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2013, pp. 1–4.
[48] R. Parhizkar, I. Dokmanic, and M. Vetterli, "Single-channel indoor microphone localization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2014, pp. 1434–1438.
[49] M. Crocco, A. Trucco, V. Murino, and A. Del Bue, "Towards fully uncalibrated room reconstruction with sound," in *Proc. Eur. Signal Process. Conf.*, Jan. 2014, pp. 910–914.
[50] I. Jager, R. Heusdens, and N. D. Gaubitch, "Room geometry estimation from acoustic echoes using graph-based echo labeling," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2016, pp. 1–5.
[51] M. Coutino, M. Bo Møller, J. Kjær Nielsen, and R. Heusdens, "Greedy alternative for room geometry estimation from acoustic echoes: A subspace-based method," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2017, pp. 366–370.
[52] 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.
[53] M. Kreißig and B. Yang, "Fast and reliable TDOA assignment in multi-source reverberant environments," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 355–359.
[54] S. Venkateswaran and U. Madhow, "Localizing multiple events using times of arrival: A parallelized, hierarchical approach to the association problem," *IEEE Trans. Signal Process.*, vol. 60, no. 10, pp. 5464–5477, Oct. 2012.
[55] 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.*, New Orleans, LA, USA, Mar. 2017, pp. 106–110.
[56] Y. A. Huang, J. Benesty, G. W. Elko, and R. M. Mersereati, "Real-time passive source localization: A practical linear-correction least-squares approach," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 8, pp. 943–956, Nov. 2001.

[57] H. C. Schau and A. Z. Robinson, "Passive source localization employing intersecting spherical surfaces from time-of-arrival differences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 35, no. 8, pp. 1223–1225, Aug. 1987.
[58] J. O. Smith and J. S. Abel, "Closed-form least-squares source location estimation from range-difference measurements," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 35, no. 12, pp. 1661–1669, Dec. 1987.
[59] Y. Chan and K. Ho, "A simple and efficient estimator for hyperbolic location," *IEEE Trans. Signal Process.*, vol. 2, no. 8, pp. 1905–1915, Aug. 1994.
[60] P. Stoica and L. Jian, "Lecture notes—Source localization from range-difference measurements," *IEEE Signal Process. Mag.*, vol. 23, no. 6, pp. 63–66, Nov. 2006.
[61] A. Beck, P. Stoica, and J. Li, "Exact and approximate solutions of source localization problems," *IEEE Trans. Signal Process.*, vol. 56, no. 5, pp. 1770–1778, May 2008.
[62] P. Svaizer, A. Brutti, and M. Omologo, "Analysis of reflected wavefronts by means of a line microphone array," in *Proc. Int. Workshop Acoust. Echo Noise Control*, 2010, pp. 1–4.
[63] P. Svaizer, A. Brutti, and M. Omologo, "Use of reflected wavefronts for acoustic source localization with a line array," in *Proc. Joint Workshop Hands-free Speech Commun. Microphone Arrays*, May 2011, pp. 165–169.
[64] T. Korhonen, "Acoustic localization using reverberation with virtual microphones," in *Proc. Int. Workshop Acoust. Echo Noise Control*, 2008, pp. 1–4.
[65] S. R. Deans, "Hough transform from the Radon transform," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-3, no. 2, pp. 185–188, Mar. 1981.
[66] D. Eberly, Distance to circles in 3D, 2015. [Online]. Available: https://www.geometrictools.com/Documentation/DistanceToCircle3.pdf. Last visited on: Jun. 8, 2017.
[67] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, 3rd ed. Cambridge, MA, USA: MIT Press, 2001.
[68] E. A. P. Habets, "Room impulse response (RIR) generator," May 2008. [Online]. Available: https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-gen erator
[69] S. Müller and P. Massarani, "Transfer-function measurement with sweeps," *J. Audio Eng. Soc.*, vol. 49, no. 6, pp. 443–471, 2001.
[70] D. Leglang, geom3d, 2009. [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/24484-geom3d

Youssef El Baba (S'17) received the M.Sc. degree in communication systems, with a specialization in signal processing, from the Swiss Federal Institute of Technology, Lausanne, Switzerland, in 2015. He is currently working toward the Doctoral degree on the topic of room geometry inference and acoustic sensing using loudspeakers. In the same year, he joined the group of E.A.P. Habets, International Audio Laboratories Erlangen, Germany. His research interests include the area of acoustic, spatial and image signal processing, as well as acoustic scene analysis and more recently electro-acoustics.

Andreas Walther received the Dipl. Ing. degree in media technology from 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 at the Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany, as a Research and Development Engineer. In 2009, he joined the Audiovisual Communication Laboratory, Swiss Federal Institute of Technology 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.

Emanuël A. P. Habets (S'02–M'07–SM'11) 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-Universität Erlangen-Nürnberg and Fraunhofer IIS), Erlangen, Germany, and the Head of the Spatial Audio Research Group with Fraunhofer IIS, Munich, Germany. From 2007 until 2009, he was a Postdoctoral Fellow with the Technion—Israel Institute of Technology and with the Bar-Ilan University, Israel. From 2009 until 2010, he was a Research Fellow with the Communication and Signal Processing Group at Imperial College London, U.K. His research interests include 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 Cochair for the 2013 International Workshop on Applications of Signal Processing to Audio and Acoustics in New Paltz, NY, USA, and the General Cochair for the 2014 International Conference on Spatial Audio (ICSA) in 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*, and an Associate Editor of the IEEE SIGNAL PROCESSING LETTERS (2013–2017). He is 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, the Vice-Chair of the EURASIP Special Area Team on Acoustic, Sound and Music Signal Processing, and the Editor-in-Chief of the *EURASIP Journal on Audio, Speech, and Music Processing*.