Proceedings of ICSA 2019

5th International Conference on Spatial Audio
September 26th to 28th, 2019, Ilmenau, Germany

DOI: 10.22032/dbt.39963

DOI (proceedings): 10.22032/dbt.39936

Full Reviewed Paper at ICSA 2019

Presented by VDT.

Room geometry inference using sources and receivers on a uniform linear array

Youssef El Baba[1], Andreas Walther[2], Emanuël A. P. Habets[3]
[1] *International Audio Laboratories Erlangen*[†]*, Germany, Email: youssef.elbaba@audiolabs-erlangen.de*
[2] *Fraunhofer Institute for Integrated Circuits, Erlangen, Germany, Email: andreas.walther@iis.fraunhofer.de*
[3] *International Audio Laboratories Erlangen*[†]*, Germany, Email: emanuel.habets@audiolabs-erlangen.de*

Abstract

State-of-the-art room geometry inference algorithms estimate the shape of a room by analyzing peaks in room impulse responses. These algorithms typically require the position of the source wrt the receiver array; this position is often estimated with sound source localization, which is susceptible to high errors under common sampling frequencies. This paper proposes a new approach, namely using an array with a known geometry and consisting of both sources and receivers. When these transducers constitute a uniform linear array, new challenges and opportunities arise for performing room geometry inference. We propose solutions designed to address these challenges, but also designed to leverage the opportunities for better results.

Keywords: Image model, time of arrival disambiguation, echo labeling, reflection point localization, reflector localization, room geometry inference.

1. 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 [1]: 3D sound analysis and reproduction, robust sound source localization (SSL), speaker tracking and de-reverberation. RGI methods use times of arrival (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. In specific, first-order TOAs characterize the physical walls present in the room. The largest family of reflector localization (RL) methods relies on ellipse geometry [2–6] or hyperbola geometry [7–9]. Other methods rely on beamforming or other schemes [1, 10]. For RL, TOAs need to be separated into sets, each set belonging to a single reflector [8]. These sets are used individually with the measurement position, either known or estimated using SSL, to define multiple constraints which together localize a reflector.

RGI can considerably benefit from a-priori knowledge of all the transducers' locations. Most importantly, finding the system latency in real measurements is a challenge [1] which can be alleviated with knowledge of the relative transducer positions. Known array geometries are commonly assumed in the RGI literature [1]; however, these usually contain either microphones or loudspeakers, exclusively. Employing an array with both types of transducers is uncommon[1]. Nonetheless, existing arrays with a single type of transducer can be transformed into arrays having both types by using one loudspeaker as a microphone or vice versa; this is made possible by acoustic transducer reciprocity. Thus, a known array

[†] A joint institution of the Friedrich-Alexander-University Erlangen-Nürnberg (FAU) and Fraunhofer IIS, Germany.

[1] Albeit there are exceptions setting a precedent for this [11].

geometry can be equivalent to known relative loudspeaker-microphone positions; this motivates our adoption of an intra-array RGI setup involving a uniform transducer array with multiple loudspeakers and one microphone (non-coincident).

This paper presents multiple adaptations to our existing RGI algorithm [1] to address the challenges of this intra-array setup, e.g., those due to shorter distances between sources and the receivers. Additionally, the paper proposes one new improvement inspired by this intra-array setup and leveraging the opportunities offered by it, as well as two more general improvements independent of the setup. The novelties and performance evaluation are presented in 2D; however, they are generalizable to 3D.

2. Room geometry inference problem and existing solution

2.1. Problem formulation

Given a uniform linear array (ULA) of L loudspeakers with a single omnidirectional microphone, and assuming that the acoustic propagation can be modeled by a linear time-invariant filter[2], the RIR of the filter between the j-th loudspeaker and the microphone (notwithstanding noise) can be expressed by

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

where α_{0j} and α_{rj} are the attenuation coefficients of the direct and reflection paths, respectively, the index r refers to one of R real or image reflectors, the function $\delta(t)$ represents the delta function and t denotes time. The TOAs τ_{0j} that arrive from the L loudspeakers to the real microphone and the TOAs τ_{rj} $(r \in \{1..R\})$ that arrive to the image microphones correspond to the direct and reflected wavefronts, respectively; they form the sets $\mathcal{T}_r = \{\tau_{rj} : \forall j \in \{1..L\}\}$.

These RIRs and the known relative positions of the loudspeakers and microphone in the array 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[3], 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 Cartesian 2D coordinate vector.

2.2. Overview of existing solution

We build upon our existing RGI method from [1], but we do not require the graph-based 3D extension it includes. This method consists of four steps. First, peaks corresponding to TOAs in the RIRs are detected and labeled using the linear Radon transform (LRT) [13]. Second, the labeled TOA sets are used to estimate the image microphone positions using [14], with knowledge of the source-receiver array geometry. Third, using the estimated image microphone positions and the array geometry, the positions of reflection points on the available reflectors are determined using the RL method in [5]. Finally, the reflection points determine the reflectors' locations and orientations. In addition to the known array geometry, this method assumes a known speed of sound and sampling frequency, which is equal across all transducers. In the case of real measurements, it also assumes zero inter-transducer latency, while allowing for a known global latency.

2.3. Challenges with intra-array setup

In this work, we assume the ULA is placed near and parallel to a reflector in the room. We use one loudspeaker in the array as a microphone; other loudspeakers are operated normally, not reciprocally. The main challenge in this setup is the near-field scenario due to the short distances between sources and receivers. This is only mitigated with lower sampling frequencies, which have the negative side effect of decreasing the precision of the LRT.

The direct sound from the nearest loudspeaker to the microphone arrives shortly after $t = 0$, and is thus disproportionally louder than the sound arriving in reflections or from farther loudspeakers; this is due to the $1/r$ sound attenuation law: in the near-field region (small distances r), differences in attenuation can be drastic between sound paths of different lengths. This causes the RIRs corresponding to the loudspeakers near the microphone to be dominated by their direct sound peaks, with reflections in these RIRs or even direct sound from other RIRs becoming relatively negligible; this translates in turn into disproportionally faint reflection responses on the LRT, especially for microphones positioned centrally on the ULA.

On the other hand, the near-field scenario violates the far-field assumption in the LRT [13]; this problem is especially noticeable for the direct sound and the wavefront reflected from the nearest wall (see Fig. 1). These wavefronts can no longer be accurately considered planar as they exhibit high curvature: the main lobes of their LRT responses are accordingly more diffuse, spread over a bigger temporal/angular region on the LRT, they attain a lower maximum amplitude and are splintered into multiple sub-responses.

3. Proposed adaptations and improvements to existing solution

3.1. Near-field adaptations

A significant contribution of this work are four adaptations designed to counter the artifacts of working in a near-field

[2] Although RIRs simulated with this image-source model [12] differ from those measured in reality, namely due to model errors, the model reproduces the early wavefronts' arrival times with sufficient accuracy for our application. This is because RGI only uses early (first- or at most second-order) reflections in rectangular rooms, and the wavefronts these produce are negligibly affected by inaccuracies of the model in simulating modal behavior or taking into account frequency-dependent absorption etc.

[3] Thus, finite reflectors are approximated by infinite planes. The final, finite room geometry can be obtained after the algorithm selects the planes corresponding to physical walls present in the room (after the region-spot-searching mode described in Section 3.2): these infinite planes intersect precisely at the boundaries of the physical walls.

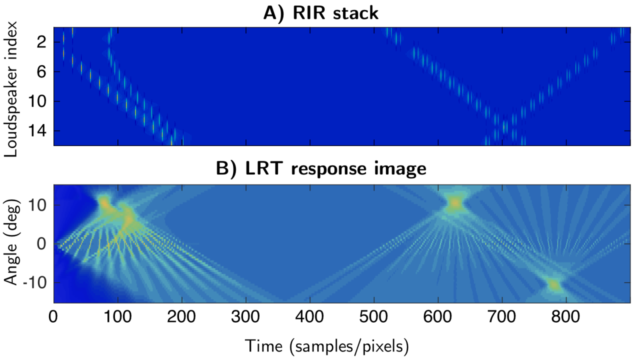

Fig. 1: Example resampled RIR stack (A) and its LRT response image (B) for a near-field scenario (Setup 1, microphone position 3 in Section 4.1, the stack is re-attenuated (see Section 3.1), and both the stack and the LRT are enhanced here for visualization). Yellow encodes high values, blue encodes low values. The two figures share the same horizontal (time) axis but have different vertical axes. Notice the lower focus of the main lobes of the LRT responses for the two earliest near-field wavefronts (in upper left region in (B)), with respect to the main lobes of the LRT responses for the later wavefronts (around samples 600-800 in (B)).

scenario.

The first adaptation selectively re-attenuates the disproportionally-boosted direct sound and earliest reflections wrt the later reflections in the RIRs, both within and outside wavefronts. Only the early region is attenuated as it is not desirable to simply compensate for the $1/r$ law for all samples: this would significantly increase noise levels in the later portion of the RIRs, with deleterious effects for the LRT; moreover, extending the re-attenuation region to later portions is also of little use since later reflected wavefronts do not suffer from near-field effects. The procedure first computes the reference direct sound TOAs $T_{j,\,\mathrm{ref}}$ for all loudspeakers $j = 1..L$ using the array geometry, then detects the earliest TOAs in each actual RIR using a peak picker; it then compares these two TOA sets to estimate any inherent global latency in the RIRs[4]. Within a temporal neighborhood T_{er} around the (latency-corrected) direct sound peaks, we re-attenuate[5] the RIRs via multiplication by r (translated into time) according to $h'_j(t) = h_j(t) f_j(t)$ with

$$f_j(t) = \begin{cases} t/(T_{\max} + T_{\mathrm{er}}/2) & \text{for } T_{1,j} \le t \le T_{2,j} \\ 1 & \text{otherwise} \end{cases}, \qquad (2)$$

and $T_{\max} = \max_{j=1..L}(T_{j,\,\mathrm{ref}})$, $T_{1,j} = \max(0, T_{j,\,\mathrm{ref}} - T_{\mathrm{er}}/2)$, $T_{2,j} = \min(TT, T_{j,\,\mathrm{ref}} + T_{\mathrm{er}}/2)$ with TT the truncated RIR length from [13]. We use $T_{\mathrm{er}} = 3a/c$ where a is the array aperture and c is the speed of sound; this T_{er} is large enough to contain the direct sound and usually also the first reflection. A side effect of this procedure is that the direct sound peaks get more similar amplitudes across all RIRs.

The second adaptation we introduce is an array-geometry-aware correction of the detected main lobe peak of the direct sound's LRT response. The procedure maps the physical angle of arrival of the direct sound wavefront to an LRT-domain angle, i.e., the angle of the incoming wavefronts on the resampled stack. For transducers on a ULA, the physical angle of arrival is always $\pm\pi/2$ (± depending on the relative ordering of the transducers), which corresponds to the maximum physically-valid angle on the LRT. However, this mapping[6] is used in more general cases (Section 3.2):

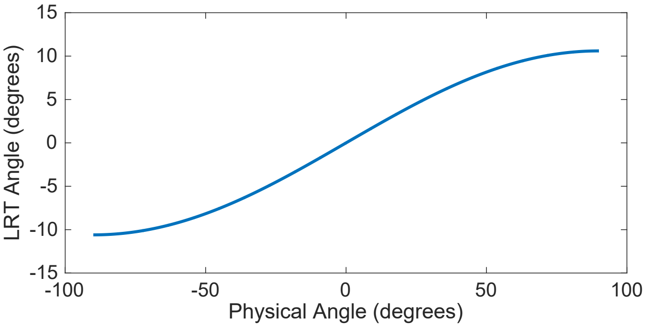

Fig. 2: Physical - LRT angle mapping from Eq.3, for $LD = 0.1$ m, $LD_{\mathrm{LRT}} = 1/750$ m, and $FS_{\mathrm{LRT}} = 48$ kHz.

$$\theta_{\mathrm{LRT}} = \mathrm{sign}(\theta_{\mathrm{phys}})\,\mathrm{atan2}\Bigg(\bigg(LD.FS_{\mathrm{LRT}}\Big/ c\Big(\big(\tan(\theta_{\mathrm{phys}} - \pi/2)\big)^2 + 1\Big)^{1/2}\bigg),\ LD/LD_{\mathrm{LRT}}\Bigg), \qquad (3)$$

where θ_{LRT} is the LRT angle and θ_{phys} is the physical angle of arrival of the wavefront to the microphone (both wrt the array center), LD is the physical transducer spacing on the array, LD_{LRT} and FS_{LRT} the transducer spacing and sampling frequency after resampling the RIR stack [13] Fig. 2 and $\mathrm{atan2}$ is the two-argument arc-tangent function. The result is then quantized to the angular grid $\mathcal{A}$ of the LRT [13] and taken as the angular bin of the main lobe. The temporal bin of the main lobe is simply given by $\sum_{j=1..L}(T_{j,\,\mathrm{ref}})/L$, and is also quantized to the sampled temporal grid. The value of the main lobe peak is then taken as the maximum LRT response inside the surrounding 7x7 region[7]. The determined LRT peak is enforced at the early stages of the processing chain; it is substituted for the LRT peaks with the 5% highest amplitudes.

The third, trivial but important adaptation is to assume the microphone position is known via the known array geometry, thereby alleviating the need for SSL.

The fourth and last adaptation addresses the neighborhood suppression size used in the LRT processing [13]. As mentioned in Section 2.3, the early wavefronts suffer from near-field effects in our setup; this translates into considerably more spurious LRT peak response detections in this region. Therefore, for short ($<$ 20 cm) minimum microphone-loudspeaker

[4] This is intended for the method to be compatible with real measurements.

[5] Strictly speaking, this is an attenuation for any distance $r < 1$ m.

[6] This mapping shares similarities with the translation formula in [13, Section 4.4], albeit going from continuous (infinite) sampling to FS_{LRT} instead of going from FS_{LRT} to FS. A more advanced version, still retaining its general shape, would use a rounding function to account for the quantized grid on the stack, however this is not considered here as the LRT used in [13] allows for further interpolation between RIR stack pixels.

[7] This region and similar parameters are chosen empirically at our resampled spatial and temporal frequencies of 750 transducers/m and 48000 kHz [13].

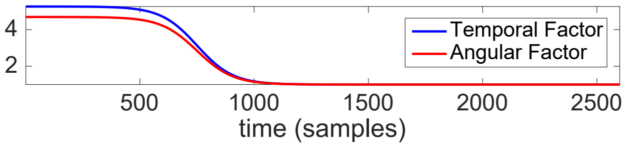

Fig. 3: Neighborhood suppression region size multiplication factors.

distances, we multiply the temporal and angular neighborhood suppression region sizes (Nl_x and Nl_y in [13], respectively) by two time-reversed sigmoid functions[8]:

$$1 + \quad 2T_{\text{er}}/3 - 1\Bigg)\Bigg(1 - 1/\Big(1 + \exp\big((T_{\text{er}} - n)/8T_{\text{er}}\big)\Big)\Bigg)\,,$$

$$1 + \quad 2|\mathcal{A}| - 1\Bigg)\Bigg(1 - 1/\Big(1 + \exp\big((T_{\text{er}} - n)/8T_{\text{er}}\big)\Big)\Bigg)\,,$$

for the temporal and angular dimensions, respectively, where n denotes the time in samples Fig. 3. The idea here is to gradually increase the suppression neighborhood going into the near-field region T_{er}. Moreover, we increase the image filter temporal width to 25, up from 15 samples (at 48 kHz) in [13], and we adjust the neighborhood suppression threshold T_r [13] to $T_r + 55\%$ for the direct-sound peak specifically[9]. Finally, in contrast to our approach in [1, 13], we also allow reflection stack-lines to intersect the direct-sound stack-lines.

3.2. Setup-inspired and general improvements

In addition to the near-field adaptations, several improvements are introduced to the RGI method. Specifically, the first two improvements are made possible by knowledge of the $\theta_{\text{phys}} - \theta_{\text{LRT}}$ angle mapping (Eq. 3).

The first improvement is the restriction, early in the processing chain, of the LRT peak response detection to physically-valid angles, i.e., angles that correspond to physical angles within $[-\pi/2, \pi/2]$. This is needed because, whereas it is not possible for sound waves to impinge on the array with bigger absolute angles, it is still theoretically possible for more-slanted but physically-invalid lines to appear on the RIR stack, and for the LRT to give a strong response to them. After all, the LRT is but a line detector in computer vision, with no such physical constraints.

The second and more significant improvement is a novel LRT region-spot-searching mode; which is promoted by the intra-array setup and which helps to achieve a more usable RL output. More specifically, the LRT peak-response detection is divided into three angular regions: 1) $\theta_{\text{phys}} \leq -40°$, 2) $-40° \leq \theta_{\text{phys}} \leq 40°$ and 3) $\theta_{\text{phys}} \geq 40°$ (all translated into LRT angles); for the lateral regions 1) and 3) we keep at most one salient LRT peak (if any), whereas for region 2) we keep the enforced (highest) direct sound peak from Section 3.1 in addition to at most the second- and third-highest peaks (if any, with a minimal time distance of a/c between these two latter). This step ensures that at most, and often exactly, four LRT peaks are detected (in addition to the direct peak); they correspond to the four walls of a rectangular room in 2D. This step solves the reflector selection problem left open (supervised) in [1, Section II-B]; it is effectively an automated reflector sifting mechanism which discards virtual (non-physical) reflectors, corresponding to second- and higher- order image microphones, and any other undesired reflector detections, e.g. the ceiling and floor detections when working in 2D with real measurements[10]. The boundaries between the regions make sense in the case of a ULA placed centrally near a wall in a shoebox room, as the image transducers corresponding to the side walls lie around $\theta_{\text{phys}} \approx \pm\pi/2$, and the image transducers corresponding to the front and back walls lie around $\theta_{\text{phys}} \approx 0$; the angular ranges of the regions are intentionally chosen broadly in order to afford an error margin for LRT peak detection and to ensure robustness to different geometrical conditions, e.g., setups where the array is placed rotated wrt – instead of parallel to – the nearby wall.

The third improvement relates to an artifact of the LRT computation when slanting the RIR stack. The LRT can theoretically detect stack-lines with *negative* central time bin when they feature an angle $|\theta_{\text{phys}}| > 0$, such as a stack-line that intersects the array-center RIR in the stack at $t = 0$ and that is rotated around this pixel. Accordingly, the LRT response is zero for $\theta_{\text{LRT}} = 0, t < 0$, but it follows a step[11] function pattern for $|\theta_{\text{LRT}}| > 0, t < 0$, especially so in the presence of noise or pre-ringing effects before the arrival of the direct sound. This step-response pattern can feign a genuine LRT response peak, especially in near-field scenarios, whereas it merely corresponds to the start of the data. Therefore, any LRT peaks within 7.5 LRT degrees and 15 samples of the artifact at $(t = 0, \theta_{\text{LRT}} = 0)$ are discarded, and any peaks with negative time bins are also discarded.

4. Performance evaluation

We perform two performance evaluations in this paper, one for TOA detection and labeling (Section 4.3) and one for RL (Section 4.4). The first evaluation gives information about how many of the reflectors are detected, whether correctly or incorrectly and how accurately (in terms of TOAs), as well as which reflectors are not detected. The second evaluation gives information about the RL error for the correctly detected physical reflectors.

4.1. Simulated setups and data sets

We re-used the same setups and performance evaluation frameworks as in [5, 13]; these consist of 7 different setups with different ULA configurations and room sizes; the only changes wrt our previous papers are the exclusion of real data and the move of the microphone positions from the cross pattern in the middle of the room (similar to [1, Fig. 11a] but in 2D) to the ULA itself, in line with the intra-array setup. To

[8] These functions and their parameters are chosen empirically.

[9] This prevents erroneously discarding the LRT peak corresponding to the image microphone of the wall near the ULA.

[10] Both of these tasks are especially challenging in setups involving arrays with limited geometrical diversity.

[11] The start of this step corresponds to the start of the data at $t = 0$ and occurs earlier for bigger absolute angles.

avoid exacerbating the already-challenging near-field effects, we only use the first three and the last three transducers on the ULA as microphones, i.e., we exclude the microphones around the array center. This means a total of $7 \cdot 6 = 42$ independent RIR stacks for testing.

4.2. Methods under test

To elucidate the impact of each set of novelties on the basis algorithm from [1], we applied different versions of the method separately on the data sets, each version including a different set of adaptations/improvements:

- Version 0: basis algorithm from [1], used[12] in 2D.
- Version 1: Version 0 with the physical angle restriction and the improvements around $t = 0$ (first and third improvements from Section 3.2).
- Version 2: Version 1 with all the near-field adaptations from Section 3.1.
- Version 3: Version 2 with the region-spot-searching mode (second improvement from Section 3.2).

We start by using our previous method unmodified from [1] (Version 0), which we then gradually but considerably expand: we first add changes independent of our intra-array setup (Version 1), then proceed to add intra-array-setup-specific adaptations (Version 2) to address the aforementioned challenges and then we finally add a major new feature to leverage the opportunities of the setup (Version 3). The distinctive advantage of Version 3 wrt Version 2 is the automatic, non-supervised discarding of second- and higher-order reflections.

4.3. TOA disambiguation metrics and results

The performance of the LRT-based TOA detection and labeling [13] was objectively assessed with three metrics: the true positive rate (TPR) indicating the percentage of detected TOA sets that match reference TOA sets, the number of false discoveries (FDs) of detected TOA sets that do not match reference TOA sets and the root mean square error (RMSE) between the correctly detected TOA sets' 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 2D simulations using the seventh-order image model [12]. Higher-order TOAs, and those beyond the truncation time *TT*, were not considered in the evaluation. All metrics were averaged across setups and microphone positions. Better performance is indicated by higher TPRs, fewer FDs and lower RMSEs. The same parameters as in [13] were used for the LRT processing.

The results (Table 1) show that the proposed adaptations result in similar robust performance as in [1]. Algorithm Versions 0 and 1 nearly fail given the intra-array setup, since they do not contain any of the adaptations addressing its challenges; this clearly motivates our proposed adaptations. More specifically, the LRT response involved in these versions is hardly usable given the aforementioned near-field effects, it only gives a splintered, diffuse response for the direct sound and a spurious artifact at $t = 0$; these two diffuse responses often temporally, and less often angularly, coincide in our intra-array setup; they disproportionally overshadow any response from reflections, and the only way to avoid this is via the near-field adaptations. In both these algorithm versions, the method can at best (albeit with difficulty) detect the direct sound properly; this explains the low TPRs. The removal of the artifact at $t = 0$ in Version 1 is inappropriate in these circumstances, as the spuriously detected artifact at $t = 0$ would itself otherwise suppress many of the spurious peaks around the genuine-but-diffuse direct-sound response; this explains the jump in the number of FDs from Version 0 to Version 1.

Tab. 1: Obtained average TOA disambiguation performance metrics.

Order	**All**	**1**		**2**	
Alg. vers.	# of FDs	TPR %	RMSE μs	TPR %	RMSE μs
0	2.12	0	N/A	0	N/A
1	3.80	20.2	363.1	1.6	96.5
2	0.55	97.0	189.5	52.0	170.5
3	0.5	97.0	191.3	0.6	145.4

Algorithm Versions 2 and 3 show remarkable and nearly-identical results (nearly-perfect first-order TPRs and a very low number of FDs). The main difference between these two versions is the nearly-complete discarding of second-order wavefront detections in Version 3; this actually fulfills the very purpose of this version: the automatic removal of second- and higher- order wavefront detections without compromising direct-sound and first-order wavefront detections (see Section 3.2).

4.4. RL metrics and results

To assess the accuracy of RL, the orientation error $O_{RL} = \left|\arccos\left(\langle \mathbf{n}_r, \widehat{\mathbf{n}}_r \rangle\right)\right|$ [15] between the true ($\mathbf{n}_r$) and estimated ($\widehat{\mathbf{n}}_r$) reflectors' normal vectors was used; additionally, the offset $D_{RL} = \left|\left|\left\langle \mathbf{n}_r, (\mathbf{m}-\mathbf{x})\right\rangle\right| - \left|\left\langle \widehat{\mathbf{n}}_r, (\mathbf{m}-\widehat{\mathbf{x}})\right\rangle\right|\right|$ [15] in terms of the distance of the true and estimated reflectors to the real microphone's true location $\mathbf{m}$ was used, where $\mathbf{x}$ and $\widehat{\mathbf{x}}$ represent points on the true and estimated reflectors, respec-

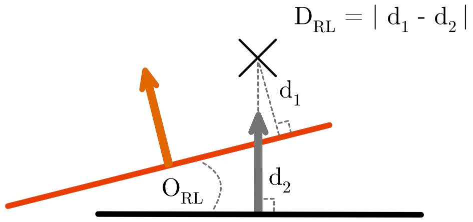

Fig. 4: Visual representation of RL error metrics. The black line, arrow and cross indicate a true reflector, its normal vector and image microphone position, while the red line and arrow indicate their estimated counterparts.

[12] We use the same parameters as [1] with the exception of the new $\widehat{R} = 10$.

Tab. 2: RL performance metrics (average values followed by ± the standard deviations), for algorithm versions 2 and 3.

Setup	Room size (m)	$\mathbf{D_{RL}}(\mathbf{cm})$	$\mathbf{O_{RL}}(^\circ)$
1	4.5x5	13.29 ± 14.13	7.45 ± 4.66
2	6x4	11.72 ± 9.20	7.66 ± 4.92
3	6x8.5	18.53 ± 26.58	7.37 ± 4.57
4	9x7.5	17.76 ± 24.17	7.32 ± 4.67
5	6x12	20.28 ± 39.26	7.35 ± 4.55
6	4.5x5	8.81 ± 14.37	6.09 ± 9.32
7	12.66x10.42	11.98 ± 8.66	3.00 ± 2.49

tively (Fig. 4). Only physical reflectors were considered, and the evaluation was done only for algorithm Versions 2 and 3; both versions gave the same metrics (Table 2), which were averaged across microphone positions (not averaged across setups). Lower metrics indicate better performance.

The results show degraded performance ($+8.04\,cm$ $\mathbf{D_{RL}}$ error and $+4.89\,^\circ$ $\mathbf{O_{RL}}$ error on average) wrt [1] (which shares identical but 3D-expanded configurations for setups 1-6) ; this is especially true for setups 4 and 5 ($+11.38/12.09\,cm$ $\mathbf{D_{RL}}$ errors and $+5.09/5.04\,^\circ$ $\mathbf{O_{RL}}$ errors, respectively). This shows that more adaptations are required to fully mitigate the near-field effects; however, it is worth noting that when the ULA is placed further away from the nearby wall and the room is larger (both conditions fulfilled in setup 7), angular error drastically decreases wrt other setups, and the distance error is also relatively lower; this is because the reflected wavefronts' near-field effects, which are not fully accounted for in the presented adaptations, are mitigated. The results are identical across algorithm versions 2 and 3 for the correctly-detected, reference-matched physical reflector detections involved; this is further evidence of the proper functioning of Version 3 (non-compromising of first-order reflections).

5. Conclusion

We presented an RGI method adapted for an intra-array transducer setup. The most important contribution in this respect is the adaptation to the near-field scenario. The second important contribution is a new mechanism for selectively sifting peak responses in the LRT domain by spot-searching in predetermined-but-broad regions; this automates the final reflector selection without compromising performance. The results show significant improvements wrt the existing RGI method from [1] with intra-array setups, with correct labeling of up to 97% of first-order echoes, albeit with degraded RL performance wrt [1] with non-intra-array setups.

6. References

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

[2] F. Antonacci, J. Filos, M. R. P. Thomas, E. A. P. Habets, A. Sarti, P. Naylor, and S. Tubaro, "Inference of room geometry from acoustic impulse responses," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 10, pp. 2683–2695, Dec. 2012.

[3] P. Annibale, J. Filos, P. Naylor, and R. Rabenstein, "Geometric inference of the room geometry under temperature variations," in *Proc. Intl. Symp. on Control, Commmunications and Signal Processing*, May 2012, pp. 1–4.

[4] 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.

[5] Y. El Baba, A. Walther, and E. A. P. Habets, "Reflector localization based on multiple reflection points," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Budapest, Hungary, Aug. 2016, pp. 1458–1462.

[6] 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.

[7] 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.

[8] 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.

[9] A. Moore, M. Brookes, and P. Naylor, "Room geometry estimation from a single channel acoustic impulse response," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Sep. 2013, pp. 1–5.

[10] I. Dokmanic, R. Parhizkar, A. Walther, Y. M. Lu, and M. Vetterli, "Acoustic echoes reveal room shape," *Proceedings of the National Academy of Sciences*, vol. 110, no. 30, pp. 12 186–12 191, 2013.

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

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

[13] Y. El Baba, A. Walther, and E. A. P. Habets, "Time of arrival disambiguation using the linear Radon transform," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 106–110.

[14] 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.

[15] J. Filos, A. Canclini, F. Antonacci, A. Sarti, and P. Naylor, “Localization of planar acoustic reflectors from the combination of linear estimates,” in *Proc. European Signal Processing Conf. (EUSIPCO)*, Aug. 2012, pp. 1019–1023.