TIME OF ARRIVAL DISAMBIGUATION USING THE LINEAR RADON TRANSFORM

Youssef El Baba[†] *Andreas Walther*[‡] *Emanuël A. P. Habets*[‡†]

[‡] Fraunhofer Institute for Integrated Circuits; Erlangen, Germany
[†]International Audio Laboratories[⋆], Erlangen, Germany

ABSTRACT

Echo labeling, the challenging task of assigning acoustic reflections to image sources, is equivalent to the highly-important disambiguation task in room geometry inference. A method using the Radon transform, an image processing tool, is proposed to address this challenge. The method relies on acoustic wavefront detection in room impulse response stacks, obtained with a uniform linear array of loudspeakers and one microphone. We show in our experiments that the proposed method can both label and detect echoes.

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

1. INTRODUCTION

Echo labeling, i.e., the task of assigning acoustic reflections to different image sources, has multiple applications: blind source separation [1], sound source localization (SSL) [2, 3] and room geometry inference (RGI) [4]. For the latter application, echo labeling refers to disambiguating which echoes — time of arrival (TOA) peaks in room impulse responses (RIRs) or time differences of arrival (TDOAs) of acoustic wavefronts — correspond to the same reflector.

Disambiguation has not been fully addressed in the RGI literature, with few exceptions [4–6], although it is a critical component. Some schemes [2, 3, 7] disambiguate TDOAs, not TOAs, which makes them vulnerable to tight TDOA error margins[1]. The authors in [2] tackle a variant of the disambiguation problem; but their algorithm is developed for SSL, not RGI. Their method is localizing multiple simultaneous sources while actively discarding reflections, rather than localizing the image sources [8] corresponding to these reflections. The authors in [3] propose improvements to the method in [2], and a re-iteration [7] re-frames the problem in terms of compatibility-conflict graphs.

Other schemes disambiguate TOAs directly. In [9], the authors rely on combinatorial candidate-source search, and pruning through Bayesian inference. The method was evaluated on TOAs obtained from simulated RIRs. The authors in [6] also propose a combinatorial scheme with validity checks, and a more efficient variant was later proposed in [10].

A drawback of [2,3,6,7,9,10] is the reliance on combinatorial search between a potentially high number of TOAs or TDOAs of reflections, which can induce computational complexity. A different approach was proposed in [4, 5] based on the generalized Hough transform (GHT), operating on detected RIR peaks. The authors in [4] move the source next to each reflector, which provides a good initial starting solution that they later refine.

In this work the linear Radon transform (LRT) is used to disambiguate TOAs. The proposed method relies on a convenient representation [11–15] of RIRs (measured with a uniform linear array of transducers) as images, referred to as stacks (Fig. 1C). The method obtains separable TOA sets in multiple steps using the available RIRs, with knowledge of the loudspeaker array geometry.

Once RIRs are visualized in stacks, distinct structures appear (Fig. 1C), corresponding to acoustic wavefronts. These arrive at the microphone from the sources and their mirrored images w.r.t. the reflectors in the room [8]. These structures can be approximated by 'stack-lines (SLs)', i.e., lines on the stack, especially when their real or image source lies far from the microphone. Therefore, detecting SLs is equivalent to obtaining a simple parametrization of the wavefronts. More specifically for TOA disambiguation, each SL corresponds to a disambiguated TOA set. Consequently, the LRT allows for detecting sets of TOAs instead of individual TOAs. These sets can then be used in estimating the locations of their corresponding real or image transducers. When the sets from all reflectors each contain TOAs for multiple loudspeakers in a *single* linear array, it becomes possible — using one of the RGI methods presented in [4, 5, 16] — to perform *2D* RGI[2].

From the perspective of a single microphone, the wavefronts can be reasonably considered planar wavefronts; namely for common shoebox room sizes, array sizes, transducer configurations and sampling frequencies. Only the early wavefronts present noticeable curvature (Fig. 1C). The later wavefronts present less curvature, due to the bigger distance between their corresponding image transducers and the array. Consequently, the approximation of wavefronts with straight lines is sufficient for all practical purposes. Moreover, based on our experiments, and strictly speaking for TOA disambiguation, SLs do not need to exactly match

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

[1]For example: with 48 kHz sampling, a TDOA from the same wavefront between a sensor pair 40 cm apart and a source 2.6 m away, takes a very small value of 4.46 samples.

[2]For 3D RGI more geometrically-diverse arrays are required.

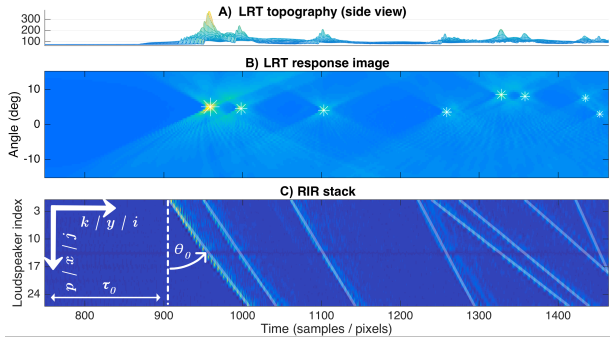

Fig. 1. The resampled RIR stack $\widetilde{\mathbf{R}}$ (C) (rescaled here for visualization), its LRT image $\mathbf{T}$ (B) and the latter's side-view topography (A). Yellow encodes high values, blue encodes low values. The stack C) is obtained from real measurements, with a microphone closer to the first loudspeakers, explaining the tilt. Faint white lines on C) indicate detected SLs, and white stars on B) indicate their corresponding LRT peaks. The coordinate system, indicated by white axes, is valid for both B) and C), these two sharing the same horizontal (time) axis but having different vertical axes.

wavefronts.

The concept of stacks already appeared in [11], where they were called 'transmission images', and lines on them were processed with supervision in [14, 15] to mitigate the deleterious effect of reverberation on SSL. The LRT is significant in its ability to detect lines, and was first discussed in the room-acoustic context in [12]. The LRT results in [12] were manually analyzed, whereas the LRT results in the method presented in this paper are automatically analyzed.

This paper is organized as follows: Section 2 formulates the problem mathematically, Section 3 introduces RIR stack images, Section 4 outlines the proposed solution, Section 5 evaluates its performance and Section 6 concludes the paper.

2. PROBLEM FORMULATION

Given a uniform linear array of L loudspeakers and a single microphone, the RIR between the microphone and the j^{th} loudspeaker can be expressed by

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

where the index r designates one of R real or image reflectors, and α_{0_j} and α_{rj} the attenuation coefficients of the direct and reflection paths, respectively. The function $\delta(t)$ represents the delta function, and $\eta_j(t)$ represents possible measurement noise. The time delays τ_{0j} and τ_{rj} represent the TOAs of the direct and reflected wavefronts, respectively. The aim is to automatically obtain from these RIRs disambiguated sets of TOAs $\{\{\tau_{rj} : \forall j \in \{1\,..\,L\}\} : \forall r \in \{0\,..\,R\}\}$, each set corresponding to a specific real or image microphone.

In the following, vectors are denoted by bold letters as in $\mathbf{v}$, with v_i referring to the i^{th} element, and matrices by bold uppercase letters as in $\mathbf{R}$, with $\mathbf{R}[j,i]$ referring to the element at the j^{th} row and i^{th} column, in the coordinate system in Fig. 1C. Calligraphic, uppercase letters as in $\mathcal{P}$ denote ordered sets. The notation $\mathbf{R}(\mathcal{P})$ refers to taking all elements in $\mathbf{R}$ indexed by a set $\mathcal{P}$. The RIRs' sampling frequency, the inter-loudspeaker distance and an RIR truncation time are denoted by *FS*, *LD* and *TT*, respectively.

3. IMAGE REPRESENTATION OF RIRs

The shared spatio-temporal structure of a set of RIRs can be emphasized when their amplitudes are stacked as illustrated in Fig. 1C. First, to remove irrelevant data, the temporal length of the RIRs is truncated to $TT \cdot FS$ samples. Second, the RIR stack is constructed as a matrix $\mathbf{R}$ with $\mathbf{R}[j,i] = |\,h_j(i)\,|$, where j is the loudspeaker index and i the sample index. Each RIR sample corresponds to a pixel on the stack. Since acoustic wavefronts produce both positive and negative amplitude deviations around the global DC value, the DC-compensated absolute RIR amplitudes are used. For RIRs measured from the loudspeakers of a uniform linear array to a single microphone, a RIR stack represents a uniform sampling of sound propagation in space and time.

Properties of the array (inter-loudspeaker distance, uniform spacing and linearity) influence the quality and usability of the obtained RIR stacks. Non-linear arrays are incompatible with this method, because for those, wavefronts on $\mathbf{R}$ violate the aforementioned approximation, whereas for non-uniformly spaced linear arrays additional computational steps can account for their non-uniformity in SL detection.

4. PROPOSED METHOD

First, $\mathbf{R}$ is resampled to a reference horizontal (temporal) and vertical (spatial) resolution, to ease the interpretation of the transform. The resampling factors are chosen such that 1 pixel on the resampled image $\widetilde{\mathbf{R}}$ corresponds to $1/FS_{\text{LRT}}$ temporally and a distance of LD_{LRT} spatially (Section 5).

4.1. Linear Radon transform

A linear Radon transform is applied to $\widetilde{\mathbf{R}}$. The resulting transformed image $\mathbf{T}$ (Fig. 1B) is given by [12]

$$\mathbf{T}[j,k] = \sum_{p}\widetilde{\mathbf{R}}[p, k + (p - \widetilde{L}/2)\cdot\tan(\theta_j))]\ , \qquad (1)$$

with $p \in \{1\,..\,\widetilde{L}\}$ denoting the resampled loudspeaker index in $\widetilde{\mathbf{R}}$. The angles θ_j are chosen in a predetermined set of angles $\mathcal{A}$, and the time indices k in $\mathcal{T} = \{1\,..\,TT \cdot FS_{\text{LRT}}\}$. Each angle θ_j in $\mathcal{A}$ produces a j^{th} row in $\mathbf{T}$ (Fig. 1C). The value of $\widetilde{\mathbf{R}}$ is interpolated at non-integer indices. Although

the LRT shares some aspects with the Ray Space transform, as described in [17], it should be noted that they are different.

The LRT can be understood as parallel sums of the stack $\widetilde{\mathbf{R}}$ along the vertical dimension, for every image rotation angle in a predetermined set $\mathcal{A}$, with different sum contributions according to the pixels' values. Since these values are actual absolute 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 in $\widetilde{\mathbf{R}}$. Hence, a distinct SL in $\widetilde{\mathbf{R}}$ translates to a strong peak in $\mathbf{T}$, for a certain angle and time bin. Conversely, each bin in $\mathbf{T}$ translates to a SL in $\widetilde{\mathbf{R}}$.

4.2. Post-processing of the transformed RIR stack

A crucial step in the proposed method is to isolate salient peaks in $\mathbf{T}$: this is achieved via image processing. Because actual wavefronts in $\widetilde{\mathbf{R}}$ only approximately correspond to SLs, many of these wavefronts produce non-punctual, spread responses in $\mathbf{T}$ (Fig. 1B). To isolate the highest peaks of these individual responses (Fig. 1A) , the following 2D image filter is used on $\mathbf{T}$:

$$\begin{pmatrix} -1 & \cdots & -1 & 0 & 1 & \cdots & 1 \\ -1 & \cdots & -1 & 0 & 1 & \cdots & 1 \\ -1 & \cdots & -1 & 0 & 1 & \cdots & 1 \end{pmatrix}, \qquad (2)$$

with each row containing W negative 1 values (Section 5), a 0 and W positive 1 values. The resulting filtered image values are then passed through the non-linearity $f(v) = \max(0, v)$. These two operations compute the smoothed horizontal image gradient and set all non-positive values to 0. This makes identifying the individual responses in $\mathbf{T}$ considerably easier, by smoothing and isolating their highest peaks.

4.3. Peak response detection and pruning

After post-processing, 2D peak picking is performed on $\mathbf{T}$. In contrast to 1D peak picking in individual RIRs, this is less error-prone due to multiple factors: the reduction by a factor L of the number of peaks to detect, the smoothing effect due to the LRT summation (Eq. 1) — discarding RIR peaks arising from individual RIR noise —, the responsivity of the LRT to lines, and image filtering. Any basic 2D peak picker (e.g., [18]) can be used to obtain a preliminary set of peak coordinates $\mathcal{P}_0$ in $\mathbf{T}$. These peaks are sorted in descending order of their amplitudes, resulting in $\mathcal{P}_1$.

Any peak in $\mathcal{P}_1$ that is located temporally earlier in $\mathbf{T}$ than the direct-sound peak[3] is discarded to obtain $\mathcal{P}_2 = \{(j,k) \in \mathcal{P}_1 : k \geq k_{\mathcal{P}_1(1)}\}$. Here, the indexing $k_{\mathcal{P}(r)}$ refers to taking the coordinate of $\mathcal{P}(r)$ along the dimension indexed by k, where $\mathcal{P}(r)$ is the r^{th} element of $\mathcal{P}$. An additional, similar constraint, which is implemented after the neighborhood suppression (discussed below), is enforcing that no detected SL intersect the direct-sound SL. Both of these constraints are due to the direct line of sight, making it impossible for any reflected wavefront to arrive earlier than the direct-sound wavefront at the microphone.

Next, a manually selected relative threshold T_r (Section 5) is used: it discards less-important peaks, such that $\mathcal{P}_3 = \{(j,k) \in \mathcal{P}_2 : \mathbf{T}(j,k) \geq T_r \cdot \mathbf{T}(\mathcal{P}_2(2))\}$. The use of T_r w.r.t. the *second* instead of the *first* most prominent peak's value is justified, because the former is almost guaranteed to correspond to a reflected wavefront's SL in $\widetilde{\mathbf{R}}$. Hence, the value of this second peak is likely to be closer to the values of peaks corresponding to SLs of other reflected wavefronts.

To ensure that i) no pair of detected SLs are too similar or duplicates and ii) few or no spurious SLs remain after the previous processing, neighborhood suppression is used for i) with a strict neighborhood of $Ns = (Ns_x, Ns_y)$ and for ii) with a more lenient neighborhood of $Nl = (Nl_x, Nl_y)$.

The suppression goes in order through each peak (j,k) in $\mathcal{P}_3$: any other peak (m,n) in $\mathcal{P}_3$ s.t.

i) $(|j-m| \leq \lfloor Ns_x/2 \rfloor) \wedge (|k-n| \leq \lfloor Ns_y/2 \rfloor)$, *or*,

ii) $\{(|j-m| \leq \lfloor Nl_x/2 \rfloor) \wedge (0 \leq k-n \leq \lfloor Nl_y/2 \rfloor)\} \wedge \{(\mathbf{T}(j,k) - \mathbf{T}(m,n))/\mathbf{T}(j,k) \geq T_r\}$,

is flagged. When for the r^{th} peak under consideration both conditions i) and ii) are checked for all other peaks in $\mathcal{P}_3$, the flagged peaks are discarded and the most prominent peak after (j,k), among the remaining ones, is processed. This sequential processing ensures an efficient and robust neighborhood suppression. A new set of peaks $\mathcal{P}_4$ is obtained in this way. It guarantees by construction that

i) no two peaks are within the same strict neighborhood Ns of each other, and

ii) also ensures that no peak has a smaller-valued neighbor (by a relative difference T_r) within an earlier neighborhood $(Nl_x, \lfloor Nl_y/2 \rfloor)$.

At this point, the method discards all SLs that intersect the direct-sound SL, resulting in $\mathcal{P}_5$. As a last step, a cap on the number of peaks is imposed, to keep only the $\widehat{R}$ most prominent peaks s.t. $\mathcal{P}_6 = \{(j_{\mathcal{P}_5(r)}, k_{\mathcal{P}_5(r)}) \in \mathcal{P}_5,\ r = 0 \,..\, \widehat{R}-1\}$.

4.4. SL translation and TOA disambiguation

To translate peaks from $\mathcal{P}_6$ into line parameters on $\mathbf{R}$, it suffices to compute for each peak r, using the r^{th} index pair $(j_{\mathcal{P}_6(r)}, k_{\mathcal{P}_6(r)})$ in $\mathcal{P}_6$, its angle and time offset: $a'_r = \tan^{-1}(FS_{\text{LRT}} \cdot LD_{\text{LRT}} \cdot \tan(\mathcal{A}(j_{\mathcal{P}_6(r)}))/(FS \cdot LD))$ and $k'_r = k_{\mathcal{P}_6(r)} \cdot FS/FS_{\text{LRT}}$, respectively.

After this parametrization, individual TOAs can be obtained by converting SL parameters to ordered sets of indices in $\mathbf{R}$. The index vectors for the r^{th} SL are $\mathbf{x}_r = [0 \,..\, L-1]^T$ and $\mathbf{y}_r = \tan(a'_r) \cdot \mathbf{x}_r + k'_r$. The pair (x_{r_j}, y_{r_j}) denotes the

[3] The highest peak-response in $\mathbf{T}$ is taken as the direct-sound peak.

individual-RIR peak indices in $\mathbf{R}$ for the j^{th} loudspeaker, and the corresponding TOA is given by $\tau_{rj} = y_{rj}/FS$.

The disambiguation is achieved by construction since each SL gives a unique set of TOAs clearly attributable to it through indexing. The disambiguated TOA sets are then simply given by $\{\{\tau_{rj} : \forall j \in \{1\,..\,L\}\} : \forall r \in \{0\,..\,\widehat{R}-1\}\}$.

5. PERFORMANCE EVALUATION

The proposed method was tested on multiple room and transducer configurations (for details please refer to [16, Section V]): 7 simulated setups with 9 independent measurement sets each, and 1 real-world setup with 8 independent measurement sets each. The multiple measurement sets corresponded to microphone positions scattered across the rooms, which had various sizes and proportions. In total, the method was evaluated on 63 simulated and 8 real RIR stacks.

The performance was objectively measured with three metrics: the true positive rate (TPR) indicating the percentage of detected SLs that map to correct reference TOA sets, the false discovery rate (FDR) indicating the percentage of detected SLs that do not map to correct reference TOA sets and the accuracy of the correctly detected SLs given by the root mean square error (RMSE) between a SL and its underlying TOAs in $\mathbf{R}$. Each detected SL was compared to 3D-simulated reference TOA sets, up to a reflection order of 7, and counted as correct when a one-to-one match with an RMSE distance of 0.5 ms or less was found. Higher-order SLs beyond the truncation time *TT* were not considered in the evaluation, whose results are shown in Table 1. The metrics were assembled separately for simulated reference TOA sets of orders 0-2, to facilitate the correct interpretation of the results. One of the six reflectors in the real measurement room (Row '7 (R)' in Table 1) was highly absorbent, producing quasi-invisible SLs in $\mathbf{R}$; thus, the 83.3% TPR for first-order SLs can be effectively considered a 100% TPR.

The parameters were chosen empirically, and were common for all evaluations. A relative threshold $T_r = 10\%$ was chosen, which behaves well with few exceptions across different sets of measured or simulated RIRs. The processing is optimized for — and run on — a resolution of $FS_{\text{LRT}} = 48$ kHz and $LD_{LRT} = 1/750$ m, chosen so that enough pixels in $\widetilde{\mathbf{R}}$ are available along the spatial dimension w.r.t. the temporal dimension. Crucially, this is a replacement for a hypothetical loudspeaker array populated with drivers LD_{LRT} apart. For this resolution and choice of $\mathcal{A} = \{-15, -14.5\,..\,14.5, 15\}$, the image filter's half-width $W = 7$ and neighborhood suppression sizes of $Ns = (10, 4)$ and $Nl = (50, 26)$ were chosen. The truncation time *TT* was set to 60 ms, for $\mathbf{R}$ to include all first-order SLs. The choice of $\widehat{R} = 15$ for $\mathcal{P}_6$ was not intended to serve as a precise pre-estimate of the number of SLs in $\mathbf{R}$, but rather as a useful cap considering the liberal threshold $T_r = 10\%$.

Order	All	0 (direct-sound)		1		2	
Setup	FDR %	TPR %	RMSE μs	TPR %	RMSE μs	TPR %	RMSE μs
1 (S)	0	100	148.2	98.2	66.4	37.7	63
2 (S)	0	100	103.3	88.9	58.3	39.5	57.6
3 (S)	1.48	100	103.3	88.9	56.1	35.2	62.5
4 (S)	0	100	74.6	96.3	62.5	35.3	54.7
5 (S)	1.48	100	103.3	79.6	148.2	36.1	62.9
6 (S)	1.48	100	143.7	98.1	67.5	35.2	65.1
7 (S)	0	100	57.2	97.9	6.1	37.8	53.4
Avg (S)	**0.63**	**100**	**104.8**	**92.6**	**66.5**	**37.7**	**59.9**
7 (R)	5.83	100	116.1	83.3	138.7	32.0	166.3

Table 1. Obtained performance metrics. '(S)' indicates simulations, '(R)' real data, and 'Avg (S)' the simulations' average.

The proposed method incorporates multiple steps to correctly identify SLs (Section 4). These are applicable because the LRT, in contrast to the GHT, takes into account the full range of RIR amplitudes. This permits easy distinction between the direct and reflected wavefronts, and more importantly, it allows for pruning SLs based on fine-grained LRT-domain values using neighborhood suppression (Section 4.3), to discard potential outliers. The procedure can be either used as a replacement for individual-RIR peak picking or as an auxiliary to it. In the latter case, the SLs obtained using the proposed method can be overlaid on the peak-picked TOAs to disambiguate these according to a nearest neighbor logic.

Due to the limited geometrical diversity of the array, the method was not able to distinguish very similar but genuinely separate SLs. This occurs for floor/ceiling or front/back wavefronts having the same direction and time of arrival for some placements of the linear array in a shoebox room. However, this limitation can be addressed by employing multiple linear arrays. Another problem, encountered when dealing with simulated RIRs, was when two such coincident SLs produced a higher response in $\mathbf{T}$ than the direct-sound SL; this was mitigated by setting all reflection coefficients to 0.5.

The evaluation shows that the proposed method performs robustly and accurately. It is not only able to disambiguate most TOA sets successfully up to second and sometimes higher order (not shown in Table 1), but also able to detect double and triple TOAs at SL intersections. The global, stack-based approach even makes it possible to fill in TOAs that are little visible in $\mathbf{R}$, as long as enough other TOAs are visible elsewhere on their corresponding SLs.

6. CONCLUSION

A novel, accurate TOA disambiguation method was presented, achieving TOA errors under 5 samples on average, and labeling of up to 98% of first-order TOAs on the evaluated data set. The method can also be used for robust TOA detection. Unlike current approaches, it avoids combinatorial search while employing an algorithm specifically designed for the problem. Future research will focus on extending the approach to make 3D RGI possible with the obtained sets.

7. REFERENCES

[1] Y. Huang, J. Benesty, and J. Chen, "A blind channel identification-based two-stage approach to separation and dereverberation of speech signals in a reverberant environment," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 882–895, Sep. 2005.

[2] 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.

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

[4] F. Antonacci, J. Filos, M. Thomas, E. 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.

[5] 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. European Signal Processing Conf. (EUSIPCO)*, Barcelona, Spain, Aug. 2011.

[6] 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.

[7] M. Kreißig and B. Yang, "Fast and reliable TDOA assignment in multi-source reverberant environments," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, May 2013, pp. 355–359.

[8] 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.

[9] 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.

[10] I. Jager, R. Heusdens, and N. D. Gaubitch, "Room geometry estimation from acoustic echoes using graph-based echo labeling," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016, pp. 1–5.

[11] A. J. Berkhout, M. M. Boone, and P. J. M. Valks, "Multichannel impulse responses for outdoor sound propagation," *J. Acoust. Soc. Am.*, vol. 98, no. 2, pp. 1169–1177, 1995.

[12] A. J. Berkhout, D. de Vries, and J. J. Sonke, "Array technology for acoustic wave field analysis in enclosures," *J. Acoust. Soc. Am.*, vol. 102, no. 5, pp. 2757–2770, 1997.

[13] D. de Vries and M. Boone, "Wave field synthesis and analysis using array technology," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 1999, pp. 15–18.

[14] P. Svaizer, A. Brutti, and M. Omologo, "Analysis of reflected wavefronts by means of a line microphone array," in *Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC)*, 2010.

[15] ——, "Use of reflected wavefronts for acoustic source localization with a line array," in *Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, May 2011, pp. 165–169.

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

[17] L. Bianchi, F. Antonacci, A. Sarti, and S. Tubaro, "The ray space transform: A new framework for wave field processing," *IEEE Trans. Signal Process.*, vol. 64, no. 21, pp. 5696–5706, Nov. 2016.

[18] C. Adrian and V. Aguilera. [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/12275-extrema-m--extrema2-m