# Spatial Sound Localization via Multipath Euclidean Distance Matrix Recovery

Mohammad J. Taghizadeh*, *Student Member, IEEE,* Afsaneh Asaei, *Member, IEEE,* Saeid Haghighatshoar, *Student Member, IEEE,* Philip N. Garner, *Senior Member, IEEE,* Hervé Bourlard, *Fellow, IEEE*

***Abstract*—A novel localization approach is proposed in order to find the position of an individual source using recordings of a single microphone in a reverberant enclosure. The multipath propagation is modeled by multiple virtual microphones as images of the actual single microphone and a multipath distance matrix is constructed whose components consist of the squared distances between the pairs of microphones (real or virtual) or the squared distances between the microphones and the source. The distances between the actual and virtual microphones are computed from the geometry of the enclosure. The microphone-source distances correspond to the support of the early reflections in the room impulse response associated with the source signal acquisition. The low-rank property of the Euclidean distance matrix is exploited to identify this correspondence. Source localization is achieved through optimizing the location of the source matching those measurements. The recording time of the microphone and generation of the source signal is asynchronous and estimated via the proposed procedure. Furthermore, a theoretically optimal joint localization and synchronization algorithm is derived by formulating the source localization as minimization of a quartic cost function. It is shown that the global minimum of the proposed cost function can be efficiently computed by converting it to a generalized trust region sub-problem. Numerical simulations on synthetic data and real data recordings obtained by practical tests show the effectiveness of the proposed approach.**

***Index Terms* – Single-channel source localization, Reverberant enclosure, Image model, Euclidean distance matrix, Synchronization, Distributed source localization.**

## I. Introduction

Sound source localization is an active area of research with applications in hands-free speech communication, virtual reality, and smart environment technologies. This task is often achieved by collection of spatial observation of multiple acoustic microphones which requires a carefully designed infrastructure. To facilitate distributed processing of ubiquitous sensory data provided by ad hoc microphone arrays, we are motivated to address the problem of single channel source localization in a reverberant enclosure.

The previous approaches to source localization are largely confined to multi-channel processing techniques. In the following, we provide a brief overview of the prior work on reverberant source localization. We investigate the feasibility of single channel localization based on the underlying concepts of multichannel techniques.

The previous studies are directed down two avenues of research: A large body of work is being conducted on variants of multi-channel filtering to estimate the time difference of arrival (TDOA) or steering the directivity pattern of a microphone array. The generalized cross-correlation is typically used where the peak location of the cross-correlation function of the signal of two microphones is mapped to an angular spectrum for direction of arrival estimation [1]. A weighting scheme is often employed to increase the robustness of this approach to noise and multi-path effect [2, 3]. The TDOA information can also be used to design a beamformer for directional sound acquisition. This procedure enables source localization by scanning the spatial space and computing the steered response power (SRP) of the microphone array for all directions; the source direction corresponds to that of maximum power. Delay-and-sum, minimum variance beamformer, as well as generalized side-lobe canceler have been the most effective techniques for source localization [4–6]. In principle, TDOA-based localization techniques rely on the correlation of multiple spatially distinct measurements of the source signal and they can not be applicable for single-channel localization.

From a different perspective, a wide range of research endeavors is dedicated to identifying and exploiting the structure underlying the localization problem; examples include subspace and spatial sparsity methods. The subspace methods exploit the rank structure of the received signals' covariance matrix and impose a stationarity assumption to accurately estimate the source location [7]. Sparse methods in the context of reverberant sound localization have been studied for model-based sparse component analysis [8–11]. It has been shown that incorporating spatial sparsity along with the underlying structure of the sparse coefficients enable super resolution in localization of simultaneous sources using very few microphones [9]. Relying on the image model for characterization of multipath propagation, this approach enables accurate localization of several simultaneous speech sources using recordings of under-determined mixtures; for instance up to eight overlapping speech sources can be localized with only four microphones. Although the principle of spatial sparsity holds for the recordings of a single microphone, it leads to ambiguities in signal reconstruction. Hence, localization can

Mohammad J. Taghizadeh (corresponding author) is with Huawei European Research Center, Munich, Germany; the present paper is written based on the work he did at Idiap Research Institute, Martigny, Switzerland. Saeid Haghighatshoar is with Technische Universität, Berlin, Germany; the present paper is written based on his work at École Polytechnique Fédérale de Lausanne, Switzerland. Afsaneh Asaei, Philip N. Garner and Hervé Bourlard are with Idiap Research Institute. Hervé Bourlard is also affiliated with École Polytechnique Fédérale de Lausanne, Switzerland. E-mails: mohammad.taghizadeh@huawei.com, saeid.haghighatshoar@tu-berlin.de, {afsaneh.asaei, phil.garner, herve.bourlard}@idiap.ch.

not be possible unless the original source signal is known *a priori*. The image model of multipath propagation, however, identifies the relation between the room impulse response and the source/microphone position. This concept is fundamental to enable single-channel localization as we shall see in the subsequent sections.

Furthermore, the data-driven learning and generative modeling of location-dependent spatial characteristics has been shown promising for sound source localization in a reverberant environment; in [12] and [13] room- and microphone location-specific models were trained on white noise signals and incorporated for 2D-localization with two microphones. Nesta and Omologo [14] presented an approach that exploited sparsity of source signals in the cross-power spectral domain and accounted in a statistical manner for deviations of the sources' spatial characteristics from an ideal anechoic propagation model caused by multipath effect.

There is very little work in single-channel sound source localization. Recent studies rely on supervised training of a model of transfer functions for various positions in the room. In [15], the authors estimate the acoustic transfer function from observed reverberant speech using a clean speech model. A maximum likelihood estimation is applied in the cepstral domain assuming a Gaussian mixture model for the source. The estimation involves two stages: in the training stage, the distant speech signal is modeled for the potential locations so the acoustic transfer function is learned. In the testing stage, the location is inferred based on the location dependent speech models.

Another supervised single-channel localization algorithm is proposed in [16]. The problem is cast as recovering the controlling parameters of linear systems using diffusion kernels. The proposed algorithm computes a diffusion kernel with a specially-tailored distance measure. The kernel integrates the local estimates of the covariance matrices of the measurements into a global structure. This structure, referred to as a manifold, enables parameterization of the measurements where the parameters represent the position of the source.

Furthermore, some methods using the (ultra-)wideband radio signals are proposed to enable single-channel localization from the initial (deterministic) support of the impulse response. In [17], the notion of virtual anchors is introduced whose locations are unknown and exploited via cooperation. Given the floor plan or the enclosure boundaries in [18], a maximum likelihood formulation of the source positioning is derived using the ranges to the virtual anchors. This approach has been shown promising, if the exact mapping between the range measurements and the reflective surfaces is known; However, no effective mechanism is devised to find the range-surface correspondences.

### A. Main Contributions and Paper Outline

In this paper, we propose a novel approach to single-channel sound source localization exploiting the information carried by spatial sound. In contrast to the previous methods, no supervised training is required. We use the image model to characterize multipath acoustic propagation. According to this model, a single microphone in a reverberant enclosure leads to virtual microphones positioned at the mirrored locations of the microphone with respect to the reflective boundary of the enclosure. A reverberant signal is a collective observation resulting from the superposition of all microphone signals. We assume that the location of the microphone is known a priori and construct a distance matrix consisting of the pairwise distances between the microphone and its images and the source. The distances between the microphone and its images are known from the geometry of the room. The distances between the source and microphones are extracted from the spikes of the room impulse response function. However, extra processing is necessary to match the spikes to their corresponding image microphones. We exploit the low-rank structure of the Euclidean distance matrix and propose a procedure for image identification while compensating for the asynchronous time offset of recording. Furthermore, a joint localization and synchronization algorithm is proposed to find the global optimum of the exploited cost function. The main contributions of this paper can be summarized as follows

- ◇ A novel approach to single-channel spatial sound localization exploiting the multipath propagation model and properties of Euclidean distance matrices.
- ◇ Algorithms to identify the virtual/real microphones from the early support (location of spikes) of the impulse response while estimating and compensating for the time offset of recording.
- ◇ Proposing a joint localization and synchronization algorithm via the global optimization of the appropriate squared range-based least square cost function.
- ◇ Extending the problem to distributed source localization framework using asynchronous recordings via aggregation of single-channel estimates.

The rest of the paper is organized as follows. The problem of source localization in a reverberant enclosure is formulated in Section II. In Section III, we explain the proposed spatial sound localization scheme based on multipath distance matrix recovery: The low-rank property of the Euclidean distance matrix (EDM) is established in Section III-A. Relying on the EDM properties, the algorithms for identifying the microphone-source distances along with localization and synchronization are devised in Section III-B. Given the microphone-source distances, a theoretically optimal method to joint localization and synchronization is proposed in Section III-C. The distributed source localization approach is elaborated in Section IV. The experimental results are presented in Section V and the conclusions are drawn in Section VI.

## II. Statement of the Problem

In this section, we set out the problem formulation and the premises underlying the proposed localization approach.

### A. Signal Model

Consider a scenario in which one microphone records the signal of an omni-directional source in a reverberant enclosure.

**TABLE I:** Summary of the notation.

| Symbol | Meaning | Symbol | Meaning |
|---|---|---|---|
| $\epsilon$ | recording time offset | $\boldsymbol{D}$ | microphone-source distance matrix; |
| $c$ | speed of sound | $D_{ij}$ | element of row $i$ and column $j$ |
| $\delta$ | delay parameter equal to $c\,\epsilon$ | $\tilde{\boldsymbol{M}}$ | microphone-source measured squared distance matrix |
| $\boldsymbol{d}$ | distance between source microphones | $\hat{\boldsymbol{M}}$ | microphone-source estimated squared distance matrix |
| $d_i$ | element $i$ of distance vector | $\boldsymbol{M}$ | microphone-source squared distance matrix |
| $N$ | number of microphones and source | $\boldsymbol{\Pi}$ | actual-virtual microphones Distance matrix |
| $R$ | number of reflectors | $\boldsymbol{X}$ | positions matrix |
| $\boldsymbol{z}$ | source location | $\hat{\boldsymbol{X}}$ | estimated positions matrix |

The single-channel observation in time domain $O(t)$ consists of two components: a filtered version of the original signal $s(t)$ convolved with impulse response of the room and an additive noise term $n(t)$, thus expressed as

$$O(t) = h(t) * s(t) + n(t). \tag{1}$$

The time domain impulse response of the enclosure is assumed to be a train of Dirac delta functions corresponding to the direct path propagation and multipath reflections stated as

$$h(t) = \sum_{r=0}^{\mathcal{T}} c_r \delta(t - \tau_r), \tag{2}$$

where $c_r$ denotes the attenuation factor of the $r^{\text{th}}$ path pertaining to the spherical propagation as well as the absorption of air and reflective surfaces; $\tau_r$ designates the delay associated with acquisition of the sound traveling the distance between the source and microphone: $\tau_0$ represents the direct path delay and $\tau_r,\ r > 0$ corresponds to the reflected signal. We denote the initial support of the room impulse response by

$$\mathit{\Lambda} = \{\tau_0, \ldots, \tau_{\mathcal{R}}\}. \tag{3}$$

The goal is to estimate the source location based on the following available prior information:

- ◇ Geometry of the room
- ◇ Location of one microphone
- ◇ Early support of room impulse response, $\mathit{\Lambda}$.

Due to asynchronous recording of signal and blind estimation of the room impulse response[1], there is an indeterminacy in support recovery so that $\tau_r = \tau_r^* + \epsilon$ where $\tau_r^*$ indicates the exact traveling time of the sound signal and $\epsilon$ is the recording time offset.

Table I summarizes the set of important notation adopted in this paper.

### B. Image Microphone Model

In this section, we introduce the notion of *virtual microphones* based on the image model of multipath propagation [20]. The image model theory asserts that a reverberant sound field generated by a single source in an enclosure can be characterized as the superposition of multiple anechoic sound fields generated by images of the source with respect to the enclosure boundaries. Thereby, the initial support of impulse response corresponds to the direct-path traveling time of multiple images of the source.

The image model as described above indicates a duality between the image of source and microphones to model the multipath propagation [21]. Indeed, the observation of the source signal in a reverberant environment can be characterized as a collective observation of multiple microphones recording the direct-path propagation of a single source. The *virtual microphone*, $m_r$ is obtained as the image of the actual microphone with respect to the $r^{\text{th}}$ reflective surface. Fig.1 illustrates this duality in modeling the multipath effect.

According to the image microphone model, $\Lambda$ is the propagation delay between the source and the set of microphones. We assume a cubic room shape in dimension $\kappa$ consisting of $R$ reflecting walls. The following relation holds between the components of $\mathit{\Lambda}$ and the distances between source and actual/virtual microphones: $\tau_r = d_r/c + \epsilon$ where $d_r$ denotes the microphone-source distance corresponding to time delay $\tau_r$; $c$ is the speed of sound and $\epsilon$ is the recording time offset.

The time delays (support) of the initial echos provide a unique signature of the room geometry [22]. As the impulse response is also a function of the source and microphone positions [20], knowing the room geometry and microphone position indicates a unique source position for a specific support structure in $\mathit{\Lambda}$[2]. The source localization thus amounts to addressing the following two problems: (1) finding the correspondence between $d_r$ and $r^{th}$ surface and (2) revealing the synchronization delay. To that end, we construct a multipath distance matrix from the pairs of microphones and source distances. The source localization is achieved exploiting the Euclidean distance matrix properties.

## III. Spatial Sound Localization

We use the low-rank structure of the Euclidean distance matrices (EDM) to develop novel source localization and synchronization algorithms. To that end, a microphone-source squared distance matrix is constructed. The actual/virtual microphones pairwise distances are assumed to be known from the prior knowledge on the room geometry. The source

[1] The room impulse response is supposed to be estimated blindly and for this reason it is subject to synchronization (and scaling) ambiguity. A method of blind room impulse response estimation based on cross-relation formula [19] is evaluated in Section V-C.

[2] The theory developed in [22] asserts that up to second order of reflections is required to gaurantee the unique map. It is evident that if the microphone is not positioned in symmetry to the walls, the unique map can be acheived. This condition is considered for the experiments presented in Section V.

distances to the microphones can be estimated from the early support of the room impulse response function. The difficulty then arises from the unknown microphone-source correspondence (mapping). Thus different distance matrices can be formed which are considered *incomplete* due to the unknown constellation of the microphone-source distance vector. Section III-A shows that the squared distance matrix has a low-rank structure. Relying on the results of Section III-A, the low-rank structure of the EDM is exploited in Section III-B to devise a method to identify the microphone-source distance vector thus referred to as EDM matrix recovery, which in turn enables estimation of the source location and synchronization. Given the microphone-source distances, a joint localization and synchronization algorithm is formulated in Section III-C as a quartic cost function whose optimal solution can be efficiently computed by converting it to an instance of generalized trust region subproblem.

### *A. Multipath Euclidean Distance Matrix Rank Deficiency*

The microphone pairwise distance matrix $\mathbf{\Pi}$ consists of components $\Pi_{ij}$ where $\Pi_{ij}$ denotes the distance between the actual/virtual microphones $i$ and $j$. These distances are assumed to be known a priori based on the image microphone model as explained in Section II-B. The vector of distances between source and actual/virtual microphones is represented as

$$\boldsymbol{d} = [d_0, \dots, d_R]^\top \tag{4}$$

where $.^\top$ denotes the transpose operator and $R$ is the number of reflectors. The microphone-source multipath distance matrix is constructed as

$$\boldsymbol{D} = \begin{bmatrix} \mathbf{\Pi} & \boldsymbol{d} \\ \boldsymbol{d}^\top & 0 \end{bmatrix}, \qquad \boldsymbol{D} \in \mathbb{R}^{N \times N} \tag{5}$$

where $N = R + 2$.

The components of $\boldsymbol{d}$ in (4) are assumed to be extracted from the identified support of the spikes in the room impulse response function $\Lambda$. Hence, $\boldsymbol{D}$ can be known after estimation of $\boldsymbol{d}$. The matrix $\boldsymbol{D}$ as formed in (5) contains zero-diagonal elements and the cross-microphone and microphone-source distances on the off-diagonals. Hence, it also has a symmetric structure.

The Euclidean distance matrix $\boldsymbol{D}$ after applying a simple transformation (Hadamard product) has low rank as stated through the following lemma.

**Lemma 1.** [23] Consider a matrix $\boldsymbol{M}_{N \times N}$ consisting of the squared pairwise distances between pairs of source and microphones embedded in $\mathbb{R}^\kappa$ defined as

$$\boldsymbol{M} = \boldsymbol{D} \circ \boldsymbol{D} = \left[D_{ij}^2\right], \quad i, j \in \{1, \dots, N\} \tag{6}$$

where $\circ$ denotes the Hadamard product. The matrix $\boldsymbol{M}$ has rank at most $\eta = \kappa + 2 < N$.

*Proof.* Let $\boldsymbol{X} \in \mathbb{R}^{\kappa \times N}$ denote the position matrix consisting of the coordinates of each node (source or microphone) and $\mathbb{1}_N$ is an all-one vector, we can write $\boldsymbol{M} = \mathbb{1}_N \mathbf{\Lambda}^\top + \mathbf{\Lambda} \mathbb{1}_N{}^\top - 2\boldsymbol{X}\boldsymbol{X}^\top$; thereby, $\boldsymbol{M}$ is the summation of three matrices where the first two of them are rank-1 and the third is rank-$\kappa$. Hence $\boldsymbol{M}$ has rank at most $\eta = \kappa + 2$. □

Based on Lemma 1, there is a strong dependency among the entries of a squared distance matrix. Recent advances in matrix recovery have shown that by exploiting the low-rank structure, $N^2$ components of $\boldsymbol{M}$ can be recovered from a subset of order $O(\eta N)$ of its entries; the mathematical demonstration of this theory is elaborated in [24] and it is not required for the purpose of this paper.

The squared distance matrix $\boldsymbol{M}$ as defined through (5)–(6) is indeterminate due to unknown permutation and offset underlying components of $\boldsymbol{d}$. This problem is addressed in the following section and the low-rank structure of $\boldsymbol{M}$ is exploited to recover the correct distances.

### *B. Multipath Euclidean Distance Matrix Recovery*

Recovery of $\boldsymbol{M}$ can be achieved through the following constrained optimization problem

$$\begin{gathered} \hat{\boldsymbol{M}} = \underset{\boldsymbol{M}}{\arg\min} \sum_{(i,j) \in E} \left( M_{ij} - \tilde{M}_{ij} \right)^2 \\ \text{subject to:} \quad \operatorname{rank}(\hat{\boldsymbol{M}}) = \eta \quad \text{and} \quad \hat{\boldsymbol{M}} \in \mathbb{EDM}^N \end{gathered} \tag{7}$$

where $E$ denote the set of indices of the measured distances and $\tilde{\boldsymbol{M}}$ is the corresponding squared distance matrix; by adopting the notation in [25], $\mathbb{EDM}^N$ refers to the convex cone of all $N \times N$ Euclidean distance matrices. Furthermore, the Euclidean distance matrix must satisfy the following properties [25]

$$\hat{\boldsymbol{M}} \in \mathbb{EDM}^N \iff \begin{cases} -\boldsymbol{a}^\top \hat{\boldsymbol{M}} \boldsymbol{a} \geq 0 \\ \mathbb{1}^\top \boldsymbol{a} = 0 \\ (\forall \|\boldsymbol{a}\| = 1) \\ \hat{\boldsymbol{M}} \in \mathbb{S}_h^N \end{cases} \tag{8}$$

for any vector $\boldsymbol{a} \in \mathbb{R}^N$, where $\mathbb{S}_h^N$ designates the space of symmetric, positive hollow matrices.

We assume that the components corresponding to the actual-virtual microphones pairwise distances $\mathbf{\Pi}$ in (6) are known based on prior knowledge on the room geometry and the actual microphone location. However, the following two problems associated with recovering $\boldsymbol{M}$ need to be resolved:

1) The correspondence between the spikes in the impulse response and the boundaries of the room.
2) The time shift for synchronization of the source signal generation and recording.

We refer to the first objective as *image identification* and to the second one as *synchronization*. These two tasks are the goal of the multipath Euclidean distance matrix recovery algorithm and are elaborated in the following sections.

*1) Image Identification:* Let $\boldsymbol{\Xi}$ denote the set of all possible permutations of the components of $\boldsymbol{d}$ defined in (4). Hence, the cardinality of $\boldsymbol{\Xi}$ is $(N-1)!$. The key idea is that for the correct permutation, the squared distance matrix (6) is low-

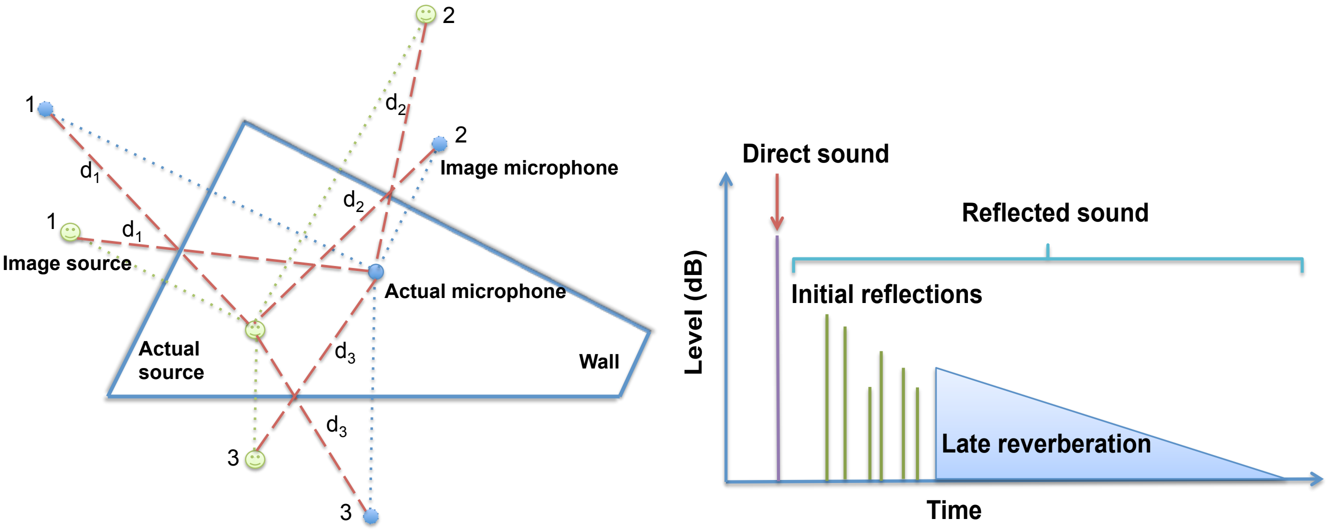

**Fig. 1:** Image microphone model of a reverberant enclosure.

rank (Lemma 1). To formalize this idea, each member $\boldsymbol{\Xi}_\pi$ of the set is used to build a distance matrix as

$$\begin{aligned} \tilde{\boldsymbol{D}}_\pi &= \begin{bmatrix} \boldsymbol{\Pi} & \boldsymbol{\Xi}_\pi \\ \boldsymbol{\Xi}_\pi^\top & 0 \end{bmatrix}, \\ \tilde{\boldsymbol{M}}_\pi &= \tilde{\boldsymbol{D}}_\pi \circ \tilde{\boldsymbol{D}}_\pi, \quad \tilde{\boldsymbol{M}}_\pi \in \mathbb{S}_h^N, \quad \pi \in [(N-1)!]\,. \end{aligned} \tag{9}$$

The goal of image identification is to suitably assign the spikes extracted from the room impulse response to their corresponding microphones. We recall that if the components of vector $\boldsymbol{\Xi}_\pi$ are in correct order, augmenting the microphone pairwise distances matrix $\boldsymbol{\Pi}$ with $\boldsymbol{\Xi}_\pi$ yields $\tilde{\boldsymbol{M}}_\pi$ of rank $\eta$.

In theory, considering only $R$ initial reflections, i.e. $N = R + 2$, seems enough to locate the first order image microphones and their correspondence to the unique location of the source. In practice, however, greater values, i.e., $\mathcal{R} > R$, can be taken into account to distinguish the echoes of the principal reflectors from the spurious peaks caused by the furniture. We discuss more on this issue in Section V-C.

*2) Synchronization:* The time discrepancy between the source signal generation and recording causes a delay of $\epsilon$ in the estimated impulse response function. In this section, we propose a new method to compensate this time shift for synchronization.

To model the effect of time difference in signal generation and recording, we define the vector $\boldsymbol{\Xi}_\pi^{\tilde{\epsilon}}$ whose $j$th element is computed as $c\,(\tau_j^\pi - \tilde{\epsilon})$ where $\tau_j^\pi$ is a member of the set $\Lambda$ (defined in (3)) at permutation $\pi$ and $\tilde{\epsilon}$ is the current estimate of $\epsilon$. Construction of $\tilde{\boldsymbol{M}}_\pi^{\tilde{\epsilon}}$ in (9) is then revised using $\boldsymbol{\Xi}_\pi^{\tilde{\epsilon}}$ for augmenting $\boldsymbol{\Pi}$ thus

$$\tilde{\boldsymbol{M}}_\pi^{\tilde{\epsilon}} = \tilde{\boldsymbol{D}}_\pi^{\tilde{\epsilon}} \circ \tilde{\boldsymbol{D}}_\pi^{\tilde{\epsilon}}, \quad \tilde{\boldsymbol{D}}_\pi^{\tilde{\epsilon}} \triangleq \tilde{\boldsymbol{D}}_\pi - \begin{bmatrix} \boldsymbol{0} & c\,\tilde{\epsilon}\,\mathbb{1}_{N-1} \\ c\,\tilde{\epsilon}\,\mathbb{1}_{N-1}^\top & 0 \end{bmatrix}. \tag{10}$$

Let us denote a vector of desired microphone-source distances with $\bar{\boldsymbol{d}}_\pi$ that corresponds to the last row of a desired squared distance matrix $\bar{\boldsymbol{M}}_\pi$. Similarly, the vector of microphone-source distances extracted from $\tilde{\boldsymbol{M}}_\pi$ is represented by $\tilde{\boldsymbol{d}}_\pi$, the synchronization delay parameter $\tilde{\delta} = c\,\tilde{\epsilon}$ is obtained through

$$\begin{aligned} &\mathcal{F}(\delta) = \left\| \bar{\boldsymbol{d}}_\pi \circ \bar{\boldsymbol{d}}_\pi - \left(\tilde{\boldsymbol{d}}_\pi - \delta\,\mathbb{1}_N\right) \circ \left(\tilde{\boldsymbol{d}}_\pi - \delta\,\mathbb{1}_N\right) \right\|_2^2 \\ &\tilde{\delta} = \arg\min_\delta \mathcal{F}(\delta) \Rightarrow \tilde{\epsilon} = \tilde{\delta}/c \end{aligned} \tag{11}$$

To solve this optimization problem, we take the derivative of the objective function $\mathcal{F}(\delta)$ and find the roots as

$$\begin{aligned} \frac{\partial \mathcal{F}(\delta)}{\partial \delta} &= -\sum_{j=1}^{N-1} 4(\tilde{d}_{\pi j} - \delta)\left((\tilde{d}_{\pi j} - \delta)^2 - \bar{d}_{\pi j}^2\right) \\ &= (N-1)\delta^3 - 3\sum_{j=1}^{N-1} \tilde{d}_{\pi j}\delta^2 + \sum_{j=1}^{N-1} (3\tilde{d}_{\pi j}^2 - \bar{d}_{\pi j}^{\,2})\delta \\ &+ \sum_{j=1}^{N-1} \left(\bar{d}_{\pi j}^{\,2}\tilde{d}_{\pi j} - \tilde{d}_{\pi j}^3\right) = 0. \end{aligned} \tag{12}$$

As the cubic polynomial in (12) has at most three roots, one can solve it analytically to find the global minimizer of the cost function defined in (11).

*3) Source Localization:* The source location is obtained from the recovered multipath distance matrix. The goal of a low-rank matrix recovery algorithm is to estimate a Euclidean distance matrix with elements as close as possible to the known entries. We use an exhaustive search through all possible permutations to solve (7) based on iterative augmentation of the distance matrix as expressed in (9). Unless $\boldsymbol{\Xi}_\pi$ consists of correct order of images, the $\tilde{\boldsymbol{M}}_\pi$ does not correspond to a Euclidean distance matrix, so we propose to project $\tilde{\boldsymbol{M}}_\pi$ on to the cone of Euclidean distance matrices, $\mathbb{EDM}^N$. To this end, we apply a projection, $\mathcal{P} : \mathbb{S}_h^N \longmapsto \mathbb{EDM}^N$ and measure the distance between the estimated matrix and the EDM cone [26, 27].

The simplest way to achieve the objective of (7) is via singular value decomposition (SVD). The projection $\mathcal{P}$ is implemented by sorting the singular values and thresholding

**Algorithm 1** SVD-Localization

```
Input: Matrix $\tilde{\boldsymbol{M}}$
Output: Estimated positions $\hat{\boldsymbol{z}}$ and synchronization delay: $\hat{\epsilon}$
1. For every $\pi \in [|\boldsymbol{\Xi}|]$ do the following steps
2. Initialize $\tilde{\epsilon} = 0$.
3. Repeat
4.     Compute $\frac{-1}{2}\boldsymbol{J}\tilde{\boldsymbol{M}}_\pi^{\tilde{\epsilon}}\boldsymbol{J}$ where $\boldsymbol{J} = \mathbb{I}_N - \frac{1}{N}\mathbb{1}_N\mathbb{1}_N^\top$
5.     Take the SVD of $\frac{-1}{2}\boldsymbol{J}\tilde{\boldsymbol{M}}_\pi^{\tilde{\epsilon}}\boldsymbol{J} = \boldsymbol{U}_\pi\boldsymbol{\Sigma}_\pi\boldsymbol{U}_\pi^\top$
6.     $\bar{\boldsymbol{X}}_\pi^{\tilde{\epsilon}} = \boldsymbol{U}_\pi^\kappa\sqrt{\boldsymbol{\Sigma}_\pi^\kappa}$, based on the largest $\kappa$ eigenvalues
7.     $\bar{\boldsymbol{\Lambda}}^{\tilde{\epsilon}} = (\bar{\boldsymbol{X}}_\pi^{\tilde{\epsilon}} \circ \bar{\boldsymbol{X}}_\pi^{\tilde{\epsilon}})\mathbb{1}_\kappa$
8.     $\bar{\boldsymbol{M}}_\pi^{\tilde{\epsilon}} = \mathbb{1}_N\bar{\boldsymbol{\Lambda}}^{\tilde{\epsilon}^\top} + \bar{\boldsymbol{\Lambda}}^{\tilde{\epsilon}}\mathbb{1}_N^\top - 2\bar{\boldsymbol{X}}_\pi^{\tilde{\epsilon}}\bar{\boldsymbol{X}}_\pi^{\tilde{\epsilon}^\top} \longrightarrow \bar{\boldsymbol{d}}_\pi$
9.     Update $\tilde{\epsilon}$ using (11).
10. Until $\tilde{\epsilon}$ converges or maximum number of iterations is reached.
11. Compute Frobenius norm of error $F_\pi = \|\tilde{\boldsymbol{M}}_\pi^{\tilde{\epsilon}} - \bar{\boldsymbol{M}}_\pi^{\tilde{\epsilon}}\|_{\mathrm{F}}$.
12. End For
13. Return Location and synchronization delay: $\hat{\epsilon}, \hat{\boldsymbol{z}} \leftarrow \arg\min_\pi F_\pi$
```

the smaller ones to achieve the desired rank. This approach is summarized in Algorithm 1. The position matrix is denoted by $\boldsymbol{X}_{N\times\kappa}$ whose $i^{\text{th}}$ row, $\boldsymbol{x}_i^\top = [x_{i1}, \dots, x_{i\kappa}], \forall i \in \{1, \dots, N-1\}$, is the position of microphone $i$ in $\kappa$-dimensional space. The order of the positions in $\boldsymbol{X}$ corresponds to the pairwise distances in $\boldsymbol{M}$. Hence, from the definition of (9), the last row corresponds to the source position $\boldsymbol{z}^\top = [z_1, \dots, z_\kappa]$.

Algorithm 1 is an *alternative coordinate descent* approach consisting of two steps. Initializing $\tilde{\epsilon}$ to zero and choosing permutation $\pi$, in the first step, the squared distance matrix $\tilde{\boldsymbol{M}}_\pi^{\tilde{\epsilon}}$ as defined in (10) is double centered [28][3] (steps 4) followed by SVD to obtain a low-rank matrix $\bar{\boldsymbol{M}}_\pi^{\tilde{\epsilon}}$ along with the position matrix $\bar{\boldsymbol{X}}_\pi^{\tilde{\epsilon}}$; $\bar{\boldsymbol{d}}_\pi$ is equal to the last row of $\bar{\boldsymbol{X}}_\pi^{\tilde{\epsilon}}$. In the second step, $\tilde{\epsilon}$ is updated by solving (11) and (12). Based on the new estimate of $\tilde{\epsilon}$, $\tilde{\boldsymbol{M}}_\pi^{\tilde{\epsilon}}$ is updated using (10). These steps are repeated until $\tilde{\epsilon}$ converges or the maximum number of iterations is reached. This procedure is executed for all possible permutations and $F_\pi = \|\tilde{\boldsymbol{M}}_\pi^{\tilde{\epsilon}} - \bar{\boldsymbol{M}}_\pi^{\tilde{\epsilon}}\|_{\mathrm{F}}$ is computed. The permutation with smallest error $F_\pi$ denotes the source location, $\hat{\boldsymbol{z}}$, and synchornisation delay $\hat{\epsilon}$.

The SVD-based low-rank projection does not incorporate the full set of EDM properties, thus it is suboptimal. More precisely, to achieve all the EDM properties, the projected matrix must satisfy the properties expressed in (8). Hence, we search in the EDM cone using the following cost function [27]

$$\mathcal{H}(\boldsymbol{X}, \tilde{\boldsymbol{M}}_\pi) = \left\| \mathbb{1}_N\boldsymbol{\Lambda}^\top + \boldsymbol{\Lambda}\mathbb{1}_N{}^\top - 2\boldsymbol{X}\boldsymbol{X}^\top - \tilde{\boldsymbol{M}}_\pi \right\|_{\mathrm{F}}^2, \tag{13}$$

where $\boldsymbol{\Lambda} = (\boldsymbol{X} \circ \boldsymbol{X})\mathbb{1}_\kappa$. The known microphone locations are used as the anchor points and only the source position is updated. The minimum of $\mathcal{H}(\boldsymbol{X}, \tilde{\boldsymbol{M}}_\pi)$ with respect to $z_i, i = \{1, \dots, \kappa\}$ can be computed by equating the partial derivative of equation (13) with respect to each individual coordinate $z_i$ to zero. Similar to (12), a third-order polynomial is obtained with maximum three roots and the one which globally minimizes the cost function is chosen. Hence, the optimization is done coordinate-wise to obtain the new estimate $\bar{\boldsymbol{X}}_\pi$ and the corresponding squared distance matrix $\bar{\boldsymbol{M}}_\pi$.

Updating $\tilde{\epsilon}$ based on (12) causes $\bar{\boldsymbol{M}}_\pi$ to deviate from the EDM cone. Hence, the optimization of (13) is repeated in an iterative fashion to project it back to the EDM cone. The stopping criterion is satisfied when the new estimate of $\tilde{\epsilon}$ differs from the old one by less than a threshold.

Although the coordinate-wise optimization procedure finds the optimal solution for each individual coordinate, reaching the global optimum is not guaranteed. Nevertheless, the experimental evaluation presented in Section V confirms that indeed the algorithm approximately converges to the optimal point. (cf. Section V).

The procedure of the EDM-Localization is summarized in Algorithm 2.

**Algorithm 2** EDM-Localization

```
Input: Matrix $\tilde{\boldsymbol{M}}$
Output: Estimated source position $\hat{\boldsymbol{z}}$ and synchronization delay: $\hat{\epsilon}$
1. For every $\pi \in [|\boldsymbol{\Xi}|]$ do the following steps do
2. Initialize $\tilde{\epsilon} = 0$.
3. Repeat
4.     $\bar{\boldsymbol{X}}_\pi^{\tilde{\epsilon}} = \arg\min_{\boldsymbol{X}} \mathcal{H}(\boldsymbol{X}, \tilde{\boldsymbol{M}}_\pi^{\tilde{\epsilon}})$
5.     $\bar{\boldsymbol{\Lambda}}^{\tilde{\epsilon}} = (\bar{\boldsymbol{X}}_\pi^{\tilde{\epsilon}} \circ \bar{\boldsymbol{X}}_\pi^{\tilde{\epsilon}})\mathbb{1}_\kappa$
6.     $\bar{\boldsymbol{M}}_\pi^{\tilde{\epsilon}} = \mathbb{1}_N\bar{\boldsymbol{\Lambda}}^{\tilde{\epsilon}^\top} + \bar{\boldsymbol{\Lambda}}^{\tilde{\epsilon}}\mathbb{1}_N^\top - 2\bar{\boldsymbol{X}}_\pi^{\tilde{\epsilon}}\bar{\boldsymbol{X}}_\pi^{\tilde{\epsilon}^\top} \longrightarrow \bar{\boldsymbol{d}}_\pi$
7.     Update $\tilde{\epsilon}$ using (11).
8. Until $\tilde{\epsilon}$ converges or maximum number of iterations is reached.
9. Compute Frobenius norm of error $F_\pi = \|\tilde{\boldsymbol{M}}_\pi^{\tilde{\epsilon}} - \bar{\boldsymbol{M}}_\pi^{\tilde{\epsilon}}\|_{\mathrm{F}}$.
10. End For
11. Return Location and synchronization delay: $\hat{\epsilon}, \hat{\boldsymbol{z}} \leftarrow \arg\min_\pi F_\pi$
```

### *C. Joint Localization and Synchronization via Generalized Trust Region Sub-problem*

In Algorithm 1 and Algorithm 2, we used an iterative approach based on low-rank SVD approximation and EDM projection to find the source location and synchronization delay. Although the resulting solutions approximate a stationary point of the cost function, there is a possibility that the resulting stationary point is a local rather than a global minimum of the cost function. Notice that since the cost function is positive and it tends to infinity as the location of the source and the synchronization delay approach infinity, the global minimum is guaranteed to exist.

In this part, we formulate finding the source location $\boldsymbol{z}$ and the delay parameter $\delta = c\epsilon$, as a quartic optimization problem. We assume that the distances between source and microphones are known based on image identification. We theoretically analyze the cost function and show that its global minimum can be efficiently computed under some mild conditions on the position of microphones and their images.

Recall that $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_{N-1}$ denote the positions of the microphones along with their images, where $\boldsymbol{x}_j \in \mathbb{R}^\kappa$ and $d_j$, $j = 1, 2, \dots, N-1$ are positive numbers corresponding to the last row of the observed square distance matrix $\tilde{\boldsymbol{M}}$ obtained after image identification; hence, $d_j$ is the measured distance between the position of the $j^{\text{th}}$ real/virtual microphone $\boldsymbol{x}_j$ and the location of the source $\boldsymbol{z}$ .

[3] Torgerson's double centering [28] as implemented in step 4 of Algorithm 1, is subtracting the row and column means of the matrix from its elements, adding the grand mean and multiplying by -1/2. The double centered matrix is scalar products relative to the origin and the coordinates is determined by the singular value decomposition (steps 5-6).

We consider the following cost function for estimating the source location $\boldsymbol{z} \in \mathbb{R}^\kappa$ and the synchronization error $\delta \in \mathbb{R}$:

$$\mathcal{G}(\boldsymbol{z}, \delta) = \sum_{j=1}^{N-1} \left( \|\boldsymbol{z} - \boldsymbol{x}_j\|^2 - (d_j - \delta)^2 \right)^2. \tag{14}$$

Let $(\hat{\boldsymbol{z}}, \hat{\delta})$ be the optimal estimate globally minimizing the cost function (14). Due to synchronization error, what is measured is a noisy distance plus some offset $\delta$ which is equal to synchronization delay multiplied with sound velocity. If this delay is compensated, the remaining noisy distances can be found by an approach similar to [29]. Hence, based on [29], we call the resulting estimate $(\hat{\boldsymbol{z}}, \hat{\delta})$ the *synchronization-extension of squared-range-based least squares* (SSR-LS) estimate. Notice that because of synchronization error, we obtain a different quartic function than [29] and as a result a completely different instance of the generalized trust region sub-problem (GTRS).

The SSR-LS cost function (14) is a non-convex quartic polynomial function of $(\boldsymbol{z}, \delta)$. Generally, it is known that global minimization of polynomials in NP-hard. However, some specific instances such as GTRS have efficient polynomial-time algorithms. In the following, we address how the global minimum of (14) can be computed efficiently.

We first transform (14) into a constrained minimization problem. Notice that

$$\mathcal{G}(\boldsymbol{z}, \delta) = \sum_{j=1}^{N-1} \left( \|\boldsymbol{z}\|^2 - \delta^2 - 2\boldsymbol{x}_j^\top \boldsymbol{z} + 2\delta d_j + \|\boldsymbol{x}_j\|^2 - d_j^2 \right)^2.$$

Therefore, setting $\gamma = \|\boldsymbol{z}\|^2 - \delta^2$, one can write

$$\min_{(\boldsymbol{z},\delta)} \mathcal{G}(\boldsymbol{z}, \delta) = \min_{(\boldsymbol{z},\delta,\gamma)} \left\{ \sum_{j=1}^{N-1} (\gamma - 2\boldsymbol{x}_j^\top \boldsymbol{z} + 2\delta d_j + \|\boldsymbol{x}_j\|^2 - d_j^2)^2 : \|\boldsymbol{z}\|^2 - \delta^2 = \gamma \right\}.$$

Assuming $\boldsymbol{y} = (\boldsymbol{z}^\top, \delta, \gamma)^\top$, this can be simplified to

$$\min_{\boldsymbol{y}} \left\{ \|\boldsymbol{A}\boldsymbol{y} - \boldsymbol{b}\|^2 : \boldsymbol{y}^\top \boldsymbol{L}\, \boldsymbol{y} + 2\boldsymbol{f}^\top \boldsymbol{y} = 0 \right\}, \tag{15}$$

where

$$\boldsymbol{A} = \begin{pmatrix} -2\boldsymbol{x}_1^\top & 2d_1 & 1 \\ -2\boldsymbol{x}_2^\top & 2d_2 & 1 \\ \vdots & \vdots & \vdots \\ -2\boldsymbol{x}_{N-1}^\top & 2d_{N-1} & 1 \end{pmatrix}, \boldsymbol{b} = \begin{pmatrix} d_1^2 - \|\boldsymbol{x}_1\|^2 \\ \vdots \\ d_{N-1}^2 - \|\boldsymbol{x}_{N-1}\|^2 \end{pmatrix}, \tag{16}$$

and

$$\boldsymbol{L} = \operatorname{diag}(\mathbb{1}_{\kappa \times 1}, -1, 0), \boldsymbol{f} = \begin{pmatrix} \boldsymbol{0}_{1\times(\kappa+1)} & -0.5 \end{pmatrix}^\top. \tag{17}$$

Matrix $\boldsymbol{A}$ has the dimension $(N-1) \times (\kappa+2)$. We assume that $N \geq (\kappa+3)$ and matrix $\boldsymbol{A}$ has full column rank which implies that $\boldsymbol{A}^\top \boldsymbol{A}$ is positive definite and, in particular, nonsingular. Note that (15) is a problem of minimizing a quadratic function under a single quadratic constraint. These kinds of problems are called *generalized trust region sub-problem* (GTRS) [30]. Although usually non-convex, GTRS problems have necessary and sufficient optimality conditions which allows them to be efficiently solved. Specially, by [30] and [29], $\boldsymbol{y} \in \mathbb{R}^{\kappa+2}$ is an optimal solution of (15) if and only if there is a $\lambda \in \mathbb{R}$ such that

$$(\boldsymbol{A}^\top \boldsymbol{A} + \lambda \boldsymbol{L})\boldsymbol{y} = \boldsymbol{A}^\top \boldsymbol{b} - \lambda \boldsymbol{f}, \tag{18}$$

$$\boldsymbol{y}^\top \boldsymbol{L}\, \boldsymbol{y} + 2\boldsymbol{f}^\top \boldsymbol{y} = 0, \tag{19}$$

$$\boldsymbol{A}^\top \boldsymbol{A} + \lambda \boldsymbol{L} \succeq \boldsymbol{0}. \tag{20}$$

Let us define

$$J_{\mathrm{PD}} = \{\lambda \in \mathbb{R} : \boldsymbol{A}^\top \boldsymbol{A} + \lambda \boldsymbol{L} \succ \boldsymbol{0}\}. \tag{21}$$

Notice that for every $\lambda \in J_{\mathrm{PD}}$, $\boldsymbol{A}^\top \boldsymbol{A} + \lambda \boldsymbol{L}$ is a positive definite thus a nonsingular matrix. We have the following useful proposition which is an application of Theorem 5.1 in [30].

**Proposition 1.** *The set $J_{PD}$ is an open interval.*

The proof is stated in the Appendix.

**Proposition 2.** *Let $J_{PD}$ be as defined in* (21). *Then, $J_{PD}$ is nonempty and bounded.*

*Proof.* We assumed that $\boldsymbol{A}$ has full column rank, which implies that $\boldsymbol{A}^\top \boldsymbol{A}$ is positive definite thus $0 \in J_{\mathrm{PD}}$. It remains to prove that $J_{\mathrm{PD}}$ is bounded from below and above.

For an upper bound, notice that $\boldsymbol{L}$ is an indefinite matrix. Let $\boldsymbol{w} = (\boldsymbol{0}_{1\times\kappa}, 1, 0)^\top$ be an all zero vector with only one 1 in position $\kappa + 1$. One can simply check that $\boldsymbol{w}^\top \boldsymbol{L}\, \boldsymbol{w} = -1$ and $\boldsymbol{w}^\top \boldsymbol{A}^\top \boldsymbol{A} \boldsymbol{w} = \|\boldsymbol{A}\boldsymbol{w}\|^2 = 4\sum_{j=1}^{N-1} d_j^2$. This implies that if $\boldsymbol{A}^\top \boldsymbol{A} + \lambda \boldsymbol{L}$ is positive definite then $\lambda < 4\sum_{j=1}^{N-1} d_j^2$. This gives an upper bound $\hat{\lambda}_u = 4\sum_{j=1}^{N-1} d_j^2$ on the interval $J_{\mathrm{PD}}$.

For a lower bound, let $\boldsymbol{v}$ be a unit norm vector with zero in its last two components. It follows that $\boldsymbol{v}^\top(\boldsymbol{A}^\top \boldsymbol{A} + \lambda \boldsymbol{L})\boldsymbol{v} > 0$ if $\lambda > -\boldsymbol{v}^\top \boldsymbol{K} \boldsymbol{v}$, where $\boldsymbol{K} = 4\sum_{j=1}^{N-1} \boldsymbol{x}_j \boldsymbol{x}_j^\top$. This implies that $\lambda > \hat{\lambda}_l = -\lambda_1(\boldsymbol{K})$, where $\lambda_1$ denotes the smallest eigenvalue of the matrix $\boldsymbol{K}$. Notice that as $\boldsymbol{A}$ is full rank, $\boldsymbol{K}$ is positive definite with $\lambda_1(\boldsymbol{K}) > 0$. Therefore $J_{\mathrm{PD}} \subset (\hat{\lambda}_l, \hat{\lambda}_u)$ and it is bounded. □

We are mostly interested in the feasible set of $\lambda$ in (18). Let us define $J_{\mathrm{PSD}} = \{\lambda \in \mathbb{R} : \boldsymbol{A}^\top \boldsymbol{A} + \lambda \boldsymbol{L} \succeq \boldsymbol{0}\}$.

**Proposition 3.** *Let $J_{PD} = (\lambda_l^*, \lambda_u^*)$ be the open interval as characterized by Proposition 2. Then, $J_{PSD} = \bar{J}_{PD} = [\lambda_l^*, \lambda_u^*]$ is a closed interval.*

*Proof.* The proof results from Theorem 5.3 in [30]. □

If we assume that the feasible $\lambda$ in (18) belongs to $J_{\mathrm{PD}}$, then $\boldsymbol{A}^\top \boldsymbol{A} + \lambda D$ is positive definite, thus one can obtain the optimal solution by

$$\hat{\boldsymbol{y}}(\lambda) = (\boldsymbol{A}^\top \boldsymbol{A} + \lambda \boldsymbol{L})^{-1}(\boldsymbol{A}^\top \boldsymbol{b} - \lambda \boldsymbol{f}). \tag{22}$$

Moreover, one can find the optimal $\lambda$ by replacing $\hat{\boldsymbol{y}}(\lambda)$ in (20) and solving the equation $\phi(\lambda) = 0, \lambda \in J_{\mathrm{PD}}$, where the function $\phi$ is defined by

$$\phi(\lambda) = \hat{\boldsymbol{y}}(\lambda)^\top \boldsymbol{L}\, \hat{\boldsymbol{y}}(\lambda) + 2\boldsymbol{f}^\top \hat{\boldsymbol{y}}(\lambda). \tag{23}$$

It is also known from [30] that $\phi(\lambda)$ is strictly decreasing over $J_{\mathrm{PD}}$. Therefore, it has only one solution which can be found

**Algorithm 3** SSR-LS

```
Input: Position of microphones and images $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_{N-1}$
Output: Estimated source position $\hat{\boldsymbol{z}}$ and synchronization delay: $\hat{\epsilon}$
1. Build $\boldsymbol{A}$, $\boldsymbol{b}$, $\boldsymbol{L}$ and $\boldsymbol{f}$ according to (16), (17)
2. Define $\hat{\boldsymbol{y}}(\lambda)$ from (22)
3. Define function $\phi(\lambda)$ from (23)
4. Set $\hat{\lambda}_u = 4\sum_{j=1}^{N-1} d_j^2$
5. Set $\hat{\lambda}_l$ to the smallest eigen-value of $\boldsymbol{K} = 4\sum_{j=1}^{N-1} \boldsymbol{x}_j \boldsymbol{x}_j^\top$
6. Solve $\phi(\lambda^*) = 0$ in the interval $(\hat{\lambda}_l, \hat{\lambda}_u)$
7. Return $(\hat{\boldsymbol{z}}, \hat{\delta})$ that is found from $\hat{\boldsymbol{y}}(\lambda^*)$
```

by applying the bisection algorithm with the initial interval estimate $(\hat{\lambda}_l, \hat{\lambda}_u)$ obtained in Proposition 2. We assume that the optimal $\lambda^*$ belongs to $J_{\text{PD}}$, thus $\boldsymbol{A}^\top \boldsymbol{A} + \lambda^* \boldsymbol{L}$ is positive definite and nonsingular. There are rare cases in which $\lambda^*$ belongs to the boundary. In our case, for example, this occurs when $\lambda^* \in \{\lambda_l^*, \lambda_u^*\}$, where $\lambda_l^*, \lambda_u^*$ are as in Proposition 3. This case, as also explained in [31], belongs to the *hard instances* of the trust region algorithm that can also be treated with a more refined analysis. In practice, considering the measurement noise, it is very rare to obtain the optimal $\lambda^*$ on the boundary.

The procedure of the proposed SSR-LS joint synchronization-localization algorithm is summarized in Algorithm 3.

An application of the proposed single-channel localization method is to devise a distributed localization framework where each microphone provides an individual estimate of the source location. The microphones may have different recording time offset which is estimated and compensated separately to yield an estimate of the source location. The single-channel estimates are then aggregated to improve the source localization performance. This idea is elaborated in the following Section IV.

## IV. Distributed Source Localization

Extension of the algorithms presented in Sections III-B and III-C to accommodate more than one microphone data is straightforward and a similar formulation as presented earlier can be applied. However, the exhaustive search required for image identification becomes prohibitive for a large network of microphones. An alternative approach is distributed source localization. That is to aggregate the individual estimates provided by each microphone while the differences in time offsets are compensated locally.

Let the estimated distances between source and microphone $l$ and its images be denoted by $\hat{\boldsymbol{d}}^l$. Furthermore, we assume that every $R+1$ consecutive rows and columns of matrix $\boldsymbol{\Pi}$ correspond to the pairwise distances between each individual microphone and its images. Hence, we can form matrix $\boldsymbol{\mathcal{D}}$ based on (5) as

$$\boldsymbol{\mathcal{D}} = \begin{bmatrix} \boldsymbol{\Pi} & [\hat{\boldsymbol{d}}^{1^\top}, \dots, \hat{\boldsymbol{d}}^{m^\top}]^\top \\ [\hat{\boldsymbol{d}}^{1^\top}, \dots, \hat{\boldsymbol{d}}^{m^\top}] & 0 \end{bmatrix}, \qquad \boldsymbol{\mathcal{D}} \in \mathbb{R}^{\mathcal{N} \times \mathcal{N}} \tag{24}$$

where $\mathcal{N} = m(R+1)+1$ and $m$ is the total number of microphones. Thereby, the squared distance matrix of the microphone array is obtained as $\boldsymbol{\mathcal{M}} = \boldsymbol{\mathcal{D}} \circ \boldsymbol{\mathcal{D}}$. As the last row of $\boldsymbol{\mathcal{M}}$ consists of the separate estimates obtained by low-rank matrix recovery performed for each microphone individually, the resulting matrix after concatenation of the distributed estimations may not fulfill the low-rank property. Thus, Algorithms 1–3 are run to yield the source location while the permutation is remained unchanged.

To summarize, the distributed microphones provide separate estimates of the microphone-source distances and the ultimate localization is achieved by estimating the source location best matching those individual estimates. The distributed localization framework can be particularly useful for ad hoc microphone setups. Further extension to multi-source scenarios is straightforward.

## V. Experimental Results

In this section, we evaluate the performance of the proposed localization algorithms using synthetic and real data recordings. We assess the robustness of the algorithms with respect to jitter noise in the support of the spikes in the room impulse response as well as synchronization delay.

### A. Single-channel Synchronization-Localization Performance

For simulation, we consider a $8 \times 5.5 \times 3.5\,\text{m}^3$ rectangular enclosure. The location of the source and microphone are randomly chosen in 100 trials. The random positions are generated such that the the distances to the boundaries are greater than 0.5 m. The speed of sound is assumed to be $c = 342\,\text{m/s}$. The sampling rate is 16 kHz.

The experiments are carried out on three simulated scenarios considering noise and synchronization delay in estimation of the room impulse response function. The noisy estimates of microphone-source distances are simulated at different noise levels indicated by distance-noise. The value of distance-noise designates the error (cm) in microphone-source distance estimation. Denoting the estimated distance from the source to microphone $j$ by $\tilde{d}_j$, we consider in our measurements $\|\tilde{d}_j - d_j\|_2 < \Delta$ and evaluation is conducted for various values of $\Delta$ as listed in the left hand side of Table II. For each scenario, we run 400 random trials and average the results. Table II summarizes the error of synchronization and source localization using SVD-Localization, EDM-Localization and SSR-LS algorithms. It is important to mention that the SVD-Localization algorithm only extracts a possibly rotated or reflected version of the points in the configuration. Using the known real/virtual microphone positions as anchor points, we use the optimization problem proposed in [32] to find the absolute position of the source whereas EDM-Localization and SSR-LS directly yield the absolute source position.

For SVD-Localization and EDM-Localization, the maximum iterations for $\hat{\epsilon}$ estimation is set to 50 and if the estimates in two successive iterations are less than 10e-5 different, the iterative synchronization is stopped earlier. It may be noted that the iterative synchronization procedure is applied only for the two first algorithms, whereas SSR-LS directly gives the $\epsilon$.

We observe that in all scenarios the image identification is achieved correctly despite the error in estimation of the microphone-source distances. Furthermore, we observe

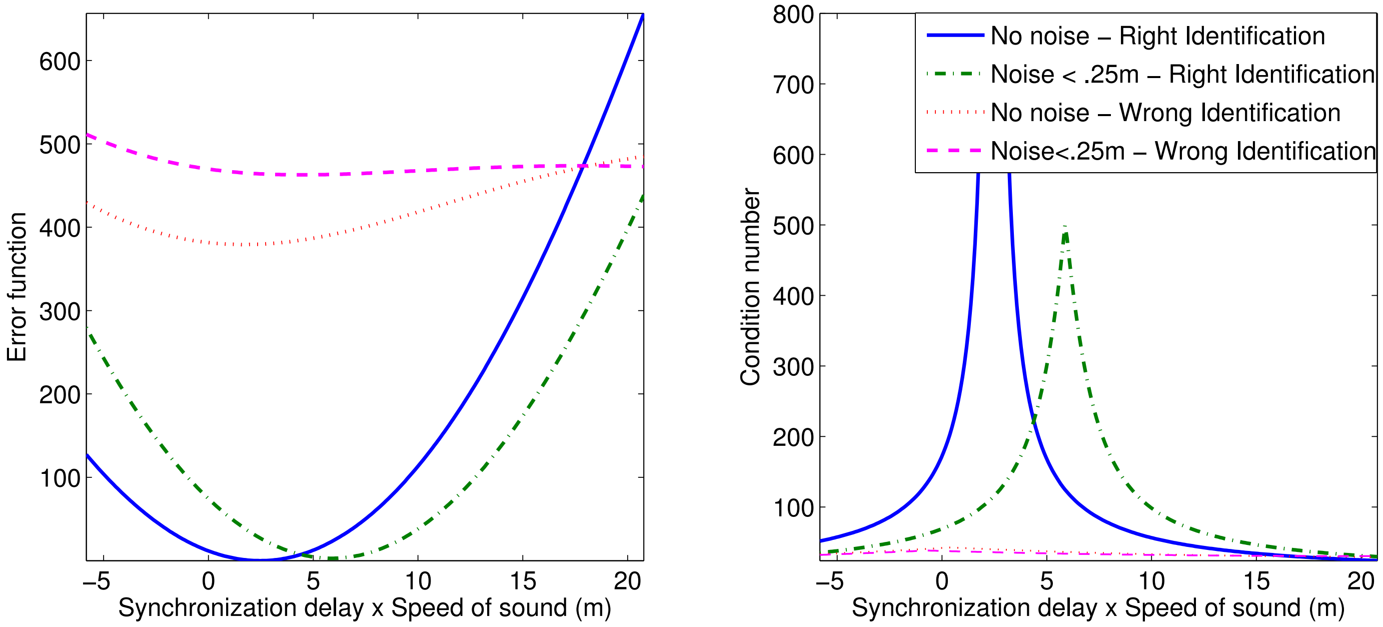

**Fig. 2:** (left) Behavior of the error function $F_\pi$ and (right) condition number of $\tilde{M}_\pi^{\tilde{\epsilon}}$ in $(10)$ for different synchronization delay when the images are identified in a right or wrong order. In this example, $\epsilon\, c = 3.4\,$m.

**TABLE II:** Performance of joint source localization and synchronization using Algorithm 1: SVD-Localization, Algorithm 2: EDM-Localization and Algorithm 3: SSR-LS Algorithm. The left hand side quantifies the level of maximum error in estimation of microphone-source distances, $\Delta$ measured in centimeters. The listed numbers quantifies the error in synchronization ($\mu$s) - finding the correct synchronization parameter $\epsilon$ - and source localization (cm) for different distance-noise levels. The numbers after ± indicate the 95% confidence interval.

| **Dis-Noise** | **Synchronization Error ($\mu$s)** | | | **Localization Error (cm)** | | |
|---|---|---|---|---|---|---|
| (cm) | **SVD-Loc.** | **EDM-Loc.** | **SSR-LS** | **SVD-Loc.** | **EDM-Loc.** | **SSR-LS** |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 23.52 ± 2.95 | 19.42 ± 1.38 | 0.71 ± 1.30 | 4.15 ± 0.12 | 3.46 ± 0.11 | 3.25 ± 0.11 |
| 10 | 44.97 ± 5.77 | 37.40 ± 2.85 | 2.49 ± 2.64 | 8.13 ± 0.25 | 6.76 ± 0.24 | 6.35 ± 0.22 |
| 15 | 70.87 ± 9.34 | 58.29 ± 4.60 | 2.96 ± 4.39 | 12.55 ± 0.37 | 10.19 ± 0.35 | 9.48 ± 0.32 |
| 20 | 100.1 ± 11.5 | 85.33 ± 6.00 | 7.14 ± 5.53 | 15.97 ± 0.52 | 13.22 ± 0.47 | 12.61 ± 0.45 |
| 25 | 123.0 ± 14.9 | 103.7 ± 8.16 | 7.61 ± 7.72 | 20.94 ± 0.66 | 16.71 ± 0.60 | 15.47 ± 0.53 |
| 30 | 148.0 ± 17.5 | 128.4 ± 9.06 | 10.09 ± 8.40 | 24.59 ± 0.74 | 20.07 ± 0.69 | 19.00 ± 0.63 |
| 40 | 204.2 ± 24.7 | 178.5 ± 12.9 | 16.94 ± 12.0 | 32.61 ± 0.99 | 27.21 ± 0.90 | 25.58 ± 0.87 |
| 50 | 242.7 ± 29.7 | 214.1 ± 16.8 | 24.84 ± 16.0 | 40.47 ± 1.27 | 33.01 ± 1.11 | 30.97 ± 1.07 |

that the results of EDM-Localization are better than SVD-Localization and they are very close to the global optimum solution of SSR-LS cost function, whereas the synchronization performance of SSR-LS is significantly better. We also observe that the the coordinate-wise minimization in the EDM-Localization almost always converges to the SSR-LS global optimum point, however, in this paper, we do not theoretically prove its global convergence.

From Table II, we see that the estimated delay parameters for the two approaches are quite different. One justification is that in the SVD method, we are looking for a three dimensional subspace as the embedding dimension for the microphone, images and the source. Now if there is a slight delay in the measurements, intuitively, this delay can be taken into account by keeping four rank-1 terms in the SVD rather than three. More precisely, the SVD method automatically takes this delay into account by adding an extra dimension for the embedding space which is removed in the truncation step in our algorithm. Intuitively that is the reason why the delay is not given exactly as in GTRS method.

Fig. 2 (left hand side) illustrates an example of the error curve for EDM-Localization Algorithm. We can see that if the augmented distance vector $\varXi_\pi$ in (9) for image identification has the correct correspondence, $\epsilon$ can be estimated with reasonable accuracy (cf. Table II). We also observe that for all the permutation except the right one, the error function $F_\pi$ has a large value and the rank of the matrix does not change much as depicted in the right hand side of Fig. 2; while the condition number of $\tilde{M}_\pi^{\tilde{\epsilon}}$ defined in (10) for the right permutation exceeds beyond 500 for noisy measurements, it is

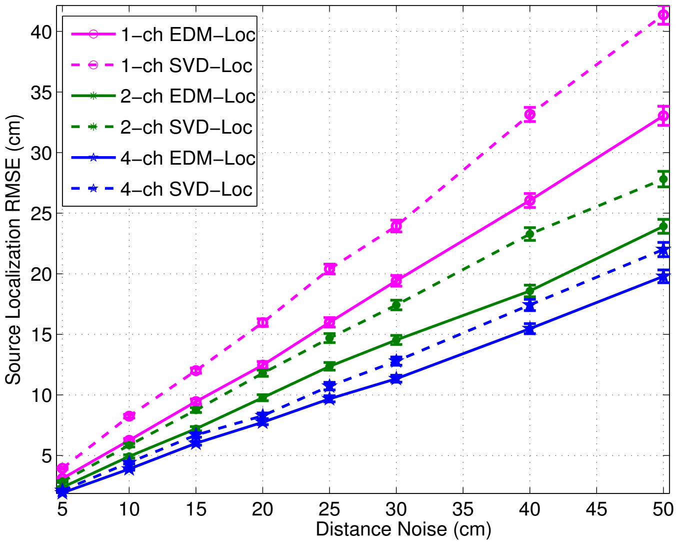

**Fig. 3:** Distributed source localization using aggregation of single microphone measurements. The error bars correspond to 95% confidence interval.

less than 25 for a wrong image identification and the measure of error is far less for a correct order. Therefore, the algorithm is able to find the correct order in all scenarios.

### *B. Multi-channel Distributed Source Localization*

The single-channel estimates can be aggregated to improve the localization performance. To that end, the microphone-source distances are estimated for each microphone and its images individually. The microphones may differ in the synchronization time offset which is estimated and compensated locally. The local distance estimates are used to construct a distance matrix as expressed in (24). Consequently, the source location will be updated using either Algorithms 1–3. The performance of the distributed source localization is illustrated in Fig. 3 for various distance-noise levels and number of microphones. The results of SSR-LS (Algorithm 3) are very close to the EDM-Localization and they are not further illustrated. The results are repeated for 100 random configurations and averaged over 400 realizations at each distance-noise level. The error bars correspond to 95% confidence interval.

We observe that exploiting additional microphones improves the source localization performance and noise robustness significantly. Furthermore, the performance gap between SVD-Localization and EDM-Localization is reduced as the number of microphones is increased. Indeed, we empirically observe that for more than ten microphones, the algorithms perform very close to each other.

### *C. Real Data Evaluation*

To conduct the real data evaluation, we use the speech recordings performed in the framework of the Multichannel Overlapping Numbers Corpus (MONC) [33]. This database was collected by outputting utterances from Numbers Corpus release 1.0 on a loudspeaker, and recording the resulting sound field using a microphone array [33] at sampling rate of 8 kHz. The recordings were made in a $8.2\,\mathrm{m} \times 3.6\,\mathrm{m} \times 2.4\,\mathrm{m}$ rectangular room containing a centrally located $4.8\,\mathrm{m} \times 1.2\,\mathrm{m}$ rectangular table. The loudspeaker was positioned at $1.2\,\mathrm{m}$ distance from the center of table at an elevation of $35\,\mathrm{cm}$ (distance from table surface to center of loudspeaker). An eight-channel, $20\,\mathrm{cm}$ diameter, circular microphone array was placed in the center of the table recorded the mixtures. The average signal to noise ratio (SNR) of the recordings was about $10\,\mathrm{dB}$. The room is mildly reverberant with a reverberation time about 250 ms.

We estimate the support of the room impulse response (RIR) function using the blind channel identification approach based on sparse cross-relation formulation [19, 34, 35]. The sparse RIR model is theoretically sound [20], and it has been shown to be useful for estimating real impulse responses in acoustic environments [36]. This approach can provide an accurate estimation of the acoustic channel up to a time delay and scaling factor. As we only need the support of the early part of the impulse response and the proposed approach can effectively handle the issue of asynchronous recording time offset, the resulting RIR is suitable to evaluate our method.

To employ the sparse cross-relation RIR estimation technique, two microphone recordings are required. Hence, the two microphones in line with the speaker ([33]) are selected. In addition to the sparsity constraint, a positivity constraint is considered to yield more accurate early support estimation [35]. The regularization parameter using the algorithm published in [19] is set to 0.3 and the CVX software package is used for optimization [37]. The length of the impulse response is set to 150 samples.

The results of the RIR estimation for one microphone are depicted in Fig. 4. The reflections are extracted by setting a threshold of 0.05 (with respect to the direct path) on the amplitude of the room impulse response. The spikes greater than this threshold define the initial support of the impulse response associated with the principal reflectors; their indices are used for distance calculation in (4). As we can see, the support is overestimated, i.e. $\mathcal{R} > R$. A single reflection (corresponding to the wall at distance 8.2 m) can not be captured in this range and thus computed based on the hypothesized correspondence at each permutation. The heuristics as such are helpful to speed-up the support recovery procedure and there is no algorithmic impediment to consider longer filters and drop the duality between the pairs of the spikes (due to parallel walls). The first spike is associated to the direct path, thus assumed to be fixed and all the $\binom{\mathcal{R}-1}{R-1}$ combinations of the support are tested. Based on the recovered support, the joint synchronization and source localization procedure estimates the source position with 5 cm error. If the measurements of two microphones are aggregated as described in Section IV, the error reduces to 3 cm.

Furthermore, we conducted experiments in more complex acoustics using the data collected at the Laboratory of Electromagnetics and Acoustics (LEMA) at École polytechnique fédérale de Lausanne (EPFL). The psychoacoustic room is considered for data collection. The dimension of the room is $6.6 \times 6.86 \times 2.69$ m$^3$ and it is fully equipped with furniture such as shelves, boxes of different textures, distributed tables and chairs. The reverberation time is about 350 ms. The source is located at $\boldsymbol{z} = [2.69 \ 1.2 \ 0.97]^\top$ with respect to the origin at the door corner. The source signal is a white Gaussian noise sampled at the rate of 51200 Hz. It is down sampled to 8000 Hz for room impulse response estimation to reduce the computational cost. If we use the channel response at a microphone located at $[2.22 \ 4.11 \ 0.95]^\top$, the source localization error is 6 cm. Using an additional microphone located at $[1.43 \, 2.71 \, 1.47]^\top$, the error is reduced to 3.5 cm.

The proposed image identification exploiting the EDM properties is robust to noise and channel order estimation and it can further be utilized to enhance the estimation of the impulse response function [35].

## VI. Conclusions

In this paper, a novel single-channel source localization approach was proposed applicable to distributed localization scenarios. The image microphone model of multipath propagation was employed to resolve the ambiguities in spatial information recovery. A multipath distance matrix was constructed where the components corresponding to the distance between the actual and virtual microphones were known. The support of the spikes in the room impulse response function indicates the distances between the unknown source location and microphones up to indeterminacies in identifying the correspondence to each image microphone along with a synchronization delay. The properties of the multipath Euclidean distance matrix were exploited to resolve these ambiguities and novel algorithms were proposed to synchronize the recordings and localize the source. In particular, an estimation strategy was derived based on globally optimizing the synchronization extension of squared-range-based least square cost function. The experiments conducted on various simulated and real data recordings of noisy scenarios demonstrated that the proposed approach is robust to jitter noise in the support of spikes in the room impulse response as well as the asynchronous time offsets in recordings. Indeed, it was shown that the synchronization delay can be estimated with reasonable accuracy and compensated for source localization. Furthermore, aggregation of multi-microphone estimates was elaborated and shown to be effective to improve the source localization performance.

## Appendix

**Proof of Proposition 1:** If $J_{\text{PD}} = \emptyset$, the argument trivially holds. Hence, let us assume that $J_{\text{PD}} \neq \emptyset$. First, we prove that $J_{\text{PD}}$ is a convex set which implies that $J_{\text{PD}}$ must be an interval. Assume that $\lambda_1, \lambda_2 \in J_{\text{PD}}$ and let $\boldsymbol{G}_i = \boldsymbol{A}^T\boldsymbol{A} + \lambda_i \boldsymbol{L}$, $i = 1, 2$. Notice that for any $\boldsymbol{u} \in \mathbb{R}^{\kappa+2}, \boldsymbol{u} \neq \boldsymbol{0}$, one has $\boldsymbol{u}^T\boldsymbol{G}_i\boldsymbol{u} > 0$, which implies that for any $\alpha \in [0, 1]$, $\boldsymbol{u}^T(\alpha\boldsymbol{G}_1 + (1-\alpha)\boldsymbol{G}_2)\boldsymbol{u} > 0$. Thus $\alpha\boldsymbol{G}_1 + (1-\alpha)\boldsymbol{G}_2 \succ \boldsymbol{0}$. As

$$\alpha\boldsymbol{G}_1 + (1-\alpha)\boldsymbol{G}_2 = \boldsymbol{A}^T\boldsymbol{A} + (\alpha\lambda_1 + (1-\alpha)\lambda_2)\boldsymbol{L},$$

it follows that for any $\alpha \in [0, 1]$, $\alpha\lambda_1 + (1-\alpha)\lambda_2 \in J_{\text{PD}}$. This proves the convexity of $J_{\text{PD}}$.

To prove the openness of the interval $J_{\text{PD}}$, let $\lambda \in J_{\text{PD}}$ be an arbitrary point and let $\boldsymbol{G} = \boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{L} \succ \boldsymbol{0}$. Consider the function $g : \{\boldsymbol{u} : \|\boldsymbol{u}\| = 1\} \to \mathbb{R}$ defined on the unit ball by $g(\boldsymbol{u}) = \boldsymbol{u}^T\boldsymbol{G}\boldsymbol{u}$. Notice that $g$ is a strictly positive function since $\boldsymbol{G} \succ \boldsymbol{0}$. Therefore, it achieves its minimum value $g^*$ on the compact set $\{\boldsymbol{u} : \|\boldsymbol{u}\| = 1\}$ where $g^* > 0$. As all the eigen-values of $\boldsymbol{L}$ consist of $\{\pm 1, 0\}$, one can simply check that $\boldsymbol{G} + \mu\boldsymbol{L} \succ \boldsymbol{0}$ for all $\mu \in (-\frac{g^*}{2}, \frac{g^*}{2})$. In particular, this implies that for all $\gamma$ in the open interval $(\lambda - \frac{g^*}{2}, \lambda + \frac{g^*}{2})$ containing $\lambda$, $\boldsymbol{A}^T\boldsymbol{A} + \gamma\boldsymbol{L} \succ 0$. This shows that $J_{\text{PD}}$ is an open interval.

## Acknowledgment

The authors would like to acknowledge Baptiste Hornecker, Hervé Lissek and Patrick Marmaroli for their support and kind help in data collection using the facilities available at the Laboratory of Electromagnetics and Acoustics (LEMA) at École Polytechnique Fédérale de Lausanne (EPFL), Switzerland.

We also would like to acknowledge the anonymous reviewers for the insightful and precise comments and remarks to improve the quality and clarity of the manuscript.

Afsaneh Asaei is supported by SNSF 200021-153507.

## References

[1] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, Signal Processing*, vol. 24 (4), pp. 320 – 327, 1976.

[2] M. Omologo and P. Svaizer, "Acoustic source localization in noisy and reverberant environments using CSP analysis," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1996.

[3] C. Blandin, A. Ozerov, and E. Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering," *Signal Processing*, vol. 92, 2012.

[4] H. T. Do, "Robust cross-correlation-based methods for sound-source localization and separation using a large-aperture microphone array," Ph.D. dissertation, Brown University in Providence, Rhode Island, United States, 2011.

[5] J. P. Dmochowski and J. Benesty, "Steered beamforming approaches for acoustic source localization," *I. Cohen, J. Benesty, and S. Gannot (Eds.), Speech Processing in Modern Communication, Springer*, vol. 24 (4), pp. 307 – 337, 2010.

[6] M. J. Taghizadeh, P. N. Garner, H. Bourlard, H. R. Abutalebi, and A. Asaei, "An integrated framework for multi-channel multi-source speaker localization and source activity detection," in *Proceedings of Hands-free Speech Communication and Microphone Arrays (HSCMA)*, 2011.

[7] J. Dmochowski, S. Benesty, and S. Affes, "Broadband music: opportunities and challenges for multiple source localization," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2007.

[8] D. Model and M. Zibulevsky, "Signal reconstruction in sensor arrays using sparse representations," *Signal Processing*, vol. 86 (3), pp. 624 – 638, 2006.

[9] A. Asaei, H. Bourlard, M. J. Taghizadeh, and V. Cevher, "Model-based sparse component analysis for reverberant speech localization," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.

[10] S. Nam and R. Gribonval, "Physics-driven structured cosparse modeling for source localization," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2012.

[11] J. L. Roux, P. T. Boufounos, K. Kang, and J. R. Hershey, "Source localization in reverberant environments using sparse optimization," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013.

[12] A. Deleforge, F. Forbes, and R. Horaud, "Variational em for binaural sound-source separation and localization," in *IEEE*

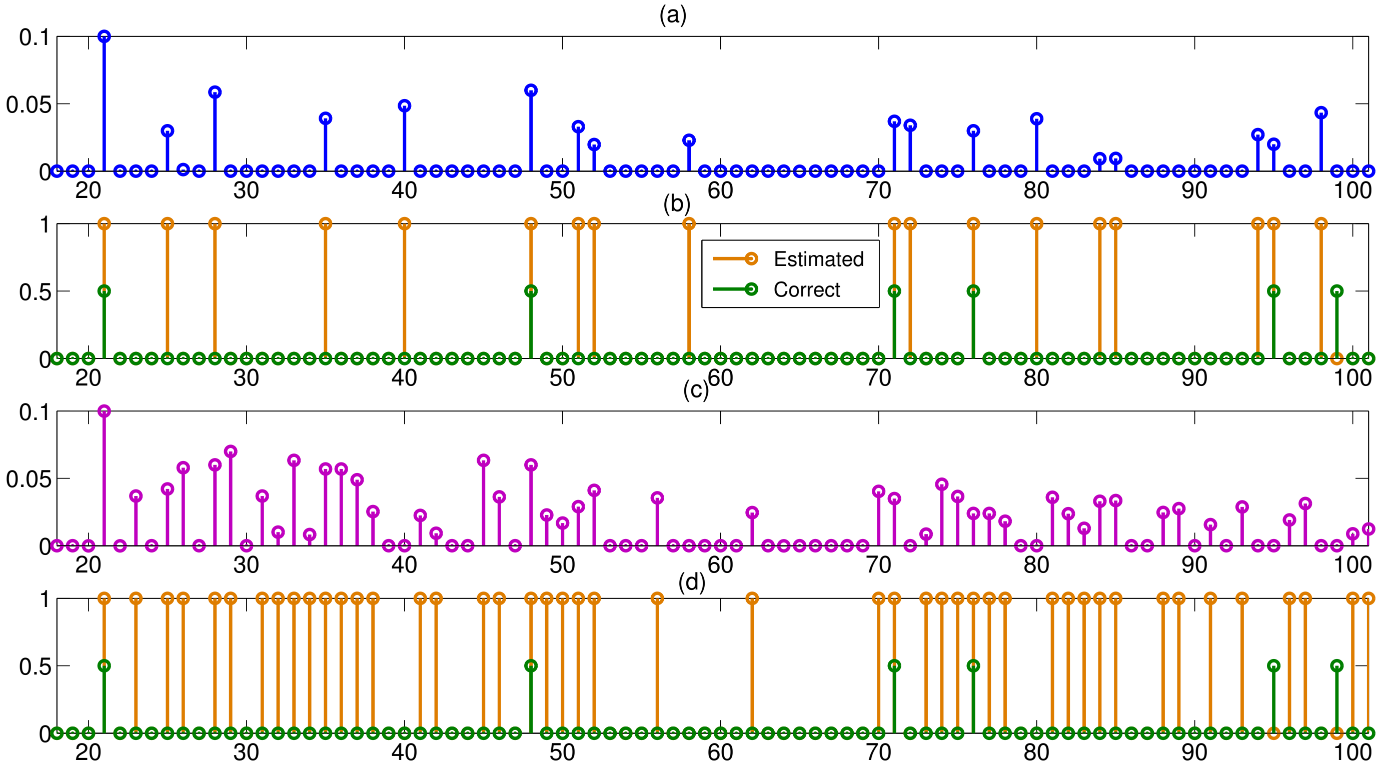

**Fig. 4:** (a) Sparse cross-relation based estimation of the early reflections in the room impulse response. (b) Support of early reflections: (orange) long stems (length=1) depicts the estimated support and the (green) short stems (length = 0.5) illustrates the true support based on the ground truth source location information. (c) Conventional cross-relation estimation of early reflections [38] and (d) Support of estimated and true early reflections: (orange) long stems (length=1) the estimated support and the (green) short stems (length = 0.5) illustrates the true support based on the ground truth source location information. Based on the estimated support of the early reflections in the room impulse response depicted in (b), the source position is estimated with $5$ cm error.

*International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 76–80.

[13] B. Laufer, R. Talmon, and S. Gannot, “Relative transfer function modeling for supervised source localization,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013, pp. 1–4.

[14] F. Nesta and M. Omologo, “Enhanced multidimensional spatial functions for unambiguous localization of multiple sparse acoustic sources,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 213–216.

[15] R. Takashima, T. Takiguchi, and Y. Ariki, *Single-Channel Sound Source Localization Based on Discrimination of Acoustic Transfer Functions*. In book: Advances in Sound Localization: InTech, 2011.

[16] R. Talmon, I. Cohen, and S. Gannot, “Supervised source localization using diffusion kernels,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2011.

[17] Y. Shen and M. Z. Win, “On the use of multipath geometry for wideband cooperative localization,” in *IEEE Global Telecommunications Conference (GLOBECOM)*, 2009.

[18] P. Meissner, C. Steiner, and K. Witrisal, “UWB positioning with virtual anchors and floor plan information,” in *IEEE 7th Workshop on Positioning Navigation and Communication (WPNC)*, 2010.

[19] Y. Lin, J. Chen, Y. Kim, and D. D. Lee, “Blind channel identification for speech dereverberation using l1-norm sparse learning,” in *Advances in Neural Information Processing Systems (NIPS)*, 2007, pp. 921–928.

[20] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small-room acoustics,” *Journal of Acoustic Society of America*, vol. 65, 1979.

[21] R. Parhizkar, I. Dokmanic, and M. Vetterli, “Single-channel indoor microphone localization,” in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. Ieee, 2014, pp. 1434–1438.

[22] I. Dokmanić, Y. M. Lu, and M. Vetterli, “Can one hear the shape of a room: The 2-D polygonal case,” in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 321–324.

[23] P. Drineas, M. Javed, M. Magdon-Ismail, G. Pandurangant, R. Virrankoski, and A. Savvides, “Distance matrix reconstruction from incomplete distance information for sensor network localization,” *Sensor and Ad Hoc Communications and Networks*, vol. 2, 2006.

[24] E. Candes and B. Recht, “Exact matrix completion via convex optimization,” *Magazine Communications of the ACM*, vol. 55, 2012.

[25] J. Dattorro, *Convex Optimization and Euclidean Distance Geometry*. USA: Meboo Publishing, 2012.

[26] M. J. Taghizadeh, A. Asaei, P. N. Garner, and H. Bourlard, “Ad hoc microphone array calibration from partial distance measurements,” in *Proceedings of Hands-free Speech Communication and Microphone Arrays (HSCMA)*, 2014.

[27] M. Taghizadeh, R. Parhizkar, P. Garner, H. Bourlard, and A. Asaei, “Ad hoc microphone array calibration: Euclidean distance matrix completion algorithm and theoretical guarantees,” *Signal Processing*, vol. 107, pp. 123–140, 2014.

[28] T. F. Cox and M. A. A. Cox, “Multidimensional scaling,”

*Chapman-Hall*, 2001.

[29] A. Beck, P. Stoica, and J. Li, "Exact and approximate solutions of source localization problems," *IEEE Transactions on Signal Processing*, vol. 56, no. 5, pp. 1770–1778, 2008.

[30] J. J. Moré, "Generalizations of the trust region problem," *Optimization Methods and Software*, vol. 2, pp. 189–209, 1993.

[31] C. Fortin and H. Wolkowicz, "The trust region subproblem and semidefinite programming," *Optimization methods and software*, vol. 19, no. 1, pp. 41–67, 2004.

[32] G. A. F. Seber, *Multivariate Observations*. Wiley & Sons, Inc, 2004.

[33] D. Moore and I. McCowan, "The multichannel overlapping numbers corpus," *Idiap resources available online:*, http://www.cslu.ogi.edu/corpora/monc.pdf.

[34] A. Aissa-El-Bey and K. Abed-Meraim, "Blind simo channel identification using a sparsity criterion," *IEEE 9th Workshop on Signal Processing Advances for Wireless Communications (SPAWC)*, pp. 271–275, 2008.

[35] A. Asaei, M. J. Taghizadeh, H. Bourlard, and V. Cevher, "Multi-party speech recovery exploiting structured sparsity models," in *12th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2011.

[36] A. Asaei, M. Golbabaee, H. Bourlard, and V. Cevher, "Structured sparsity models for reverberant speech separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 3, pp. 620–633, 2014.

[37] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 1.21," 2011, http://cvxr.com/cvx.

[38] G. Xu, H. Liu, L. Tong, and T. Kailath, "A least-squares approach to blind channel identification," *IEEE Transactions on Signal Processing*, vol. 43, pp. 2982–2993, 1995.

**Mohammad Javad Taghizadeh** received his PhD from École Polytechnique Fédérale de Lausanne, Switzerland in 2015. During his doctoral studies he was a research assistant at Idiap Research Institute in Martigny, Switzerland where he conducted research on ad hoc microphone arrays signal processing techniques to enable higher level distant speech applications. Prior to doctoral studies, he was research engineer at Information Processing group at École Polytechnique Fédérale de Lausanne and Telecommunication research centres in Tehran, Iran. He is currently a senior researcher on sound field analysis and synthesis for 3D audio technologies at Huawei European Research Centre in Munich, Germany.

**Afsaneh Asaei** received the B.S. degree from Amirkabir University of Technology and the M.S. (honors) degree from Sharif University of Technology, in Electrical and Computer engineering, respectively. She held a research engineer position at Iran Telecommunication Research Center (ITRC) from 2002 to 2008. She then joined Idiap Research Institute in Martigny, Switzerland, and was a Marie Curie fellow on speech communication with adaptive learning training network. She received the Ph.D. degree in 2013 from École Polytechnique Fédérale de Lausanne. Her thesis focused on model-based sparsity for reverberant speech processing, and its key idea was awarded the IEEE Spoken Language Processing Grant. Currently, she is a postdoctoral researcher at Idiap Research Institute. She has been a member of the technical program committee of the 4th Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA 2014) and severed as a referee of numerous journal and conference publications. Currently, she is a guest editor of Speech Communication special issue on Advances in Sparse Modeling and Low-rank Modeling for Speech Processing. Her research interests lie in the areas of signal processing, machine learning, statistics, acoustics, auditory scene analysis and cognition, and sparse signal recovery and acquisition.

**Saeid Haghighatshoar** (S'12) received the B.Sc. degree in electrical engineering in 2007 in Electronics and the M.Sc. degree in electrical engineering in 2009 in Communication Systems both from Sharif University of Technology, Tehran, Iran, and the Ph.D. degree in Computer and Communication Sciences from École Polytechnique Fédérale de Lausanne (EPFL), Switzerland. His research interests lie in information theory, Communication systems, Graphical models and Compressed sensing.

**Phil Garner** received the degree of M.Eng. in Electronic Engineering from the University of Southampton, U.K., in 1991, and the degree of Ph.D. (by publication) from the University of East Anglia, U.K., in 2012. He first joined the Royal Signals and Radar Establishment in Malvern, Worcestershire working on pattern recognition and later speech processing. In 1998 he moved to Canon Research Centre Europe in Guildford, Surrey, where he designed speech recognition metadata for retrieval. In 2001, he was seconded (and subsequently transfered) to the speech group at Canon Inc. in Tokyo, Japan, to work on multilingual aspects of speech recognition and noise robustness. As of April 2007, he is a senior research scientist at Idiap Research Institute, Martigny, Switzerland, where he continues to work in research and development of speech recognition, synthesis and signal processing. He is a senior member of the IEEE, and has published internationally in conference proceedings, patent, journal and book form as well as serving as coordinating editor of ISO/IEC 15938-4 (MPEG-7 Audio).

**Hervé Bourlard** received the Electrical and Computer Science Engineering degree and the PhD degree in Applied Sciences both from "Faculté Polytechnique de Mons", Mons, Belgium. After having been a member of the Scientific Staff at the Philips Research Laboratory of Brussels and an R&D Manager at L&H SpeechProducts, he is now Director of the Idiap Research Institute, Full Professor at the Swiss Federal Institute of Technology Lausanne (EPFL), and Founding Director of the Swiss NSF National Centre of Competence in Research on "Interactive Multimodal Information Management (IM2)". Having spent (since 1988) several long-term and short-term visits (initially as a Guest Scientist) at the International Computer Science Institute (ICSI), Berkeley, CA, he is now a member of an ICSI External Fellow and a member of its Board of Trustees.

His main research interests mainly include statistical pattern classification, signal processing, multi-channel processing, artificial neural networks, and applied mathematics, with applications to a wide range of Information and Communication Technologies, including spoken language processing, speech and speaker recognition, language modeling, multimodal interaction, augmented multi-party interaction, and distant group collaborative environments.

H. Bourlard is the author/coauthor/editor of 8 books and over 330 reviewed papers (including one IEEE paper award) and book chapters. He is a Fellow of IEEE and ISCA and a Senior Member of ACM. He is (or has been) a member of the program/scientific committees of numerous international conferences (e.g., General Chairman of IEEE Workshop on Neural Networks for Signal Processing 2002, Co-Technical Chairman of IEEE ICASSP 2002, General Chairman of Interspeech 2003) and on the Editorial Board of several journals (e.g., past co-Editor-in-Chief of "Speech Communication"). He is the recipient of several scientific and entrepreneurship awards.