HOW TO LOCALIZE TEN MICROPHONES IN ONE FINGER SNAP

Ivan Dokmanić[†], *Laurent Daudet*[‡] *and Martin Vetterli*[†]

[†]School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne (EPFL),
CH-1015 Lausanne, Switzerland
{ivan.dokmanic, martin.vetterli}@epfl.ch

[‡]Institut Langevin
ESPCI, Université Paris Diderot,
10 Rue Vauquelin 75005 Paris, France
laurent.daudet@espci.fr

ABSTRACT

A compelling method to calibrate the positions of microphones in an array is with sources at unknown locations. Remarkably, it is possible to reconstruct the locations of both the sources and the receivers, if their number is larger than some prescribed minimum [1, 2]. Existing methods, based on times of arrival or time differences of arrival, only exploit the direct paths between the sources and the receivers. In this proof-of-concept paper, we observe that by placing the whole setup inside a room, we can reduce the number of sources required for calibration. Moreover, our technique allows us to compute the absolute position of the microphone array in the room, as opposed to knowing it up to a rigid transformation or reflection. The key observation is that echoes correspond to virtual sources that we get "for free". This enables endeavors such as calibrating the array using only a single source.

Index Terms— Localization, array calibration, indoor calibration, echo sorting, microphone array

1. INTRODUCTION

Most applications of microphone arrays require us to know their relative positions. We are interested in two big groups of applications. First group is ad-hoc microphone arrays. This could be microphones that are deployed to run an experiment or make a recording, or an array of microphones on devices that happen to share the room (smartphones, tablets, laptops, glasses). Another relevant group of applications is in very large, fixed microphone arrays, where precise hand measurements of the microphone positions become impossible. By very large we think of at least several tens, or even hundreds or thousands of microphones [3].

In the recent years, a number of methods were developed for automatic position calibration of ad-hoc microphone arrays. These methods replace slow, imprecise and impractical hand measurements. Some methods use specialized devices, such as loudspeakers mounted on a fixed construction [4], or assume partial knowledge of the array geometry.

An interesting class of methods perform the calibration with sources at unknown positions. Authors in [5] formulate a non-linear least squares problem using at least five loudspeakers, and derive a closed form solution in the case when one loudspeaker is close to a microphone. In [6], the authors demonstrate an approach that uses low-rank matrix factorization. When one microphone and one source are collocated, they too derive a closed-form expression for microphone positions. With sources appearing at known times (as the authors assume), the problem reduces to multi-dimensional unfolding, and the solution is similar to that in [7].

Some methods can additionally handle unknown offset times. Thrun [8] reported a matrix factorization based method that assumes the sources to be in the far-field. No far-field assumption is exhibited in [1] and [2]. In [2] the authors also allow for unknown internal delays of the microphone processing chain. Not knowing when the sources appeared prohibits us from directly accessing the absolute distances between the sources and the microphones.

In the existing approaches, the room reverberation is either ignored or considered detrimental. We demonstrate that being in a room facilitates the calibration, even if we do not know how the room looks like or how the microphone array is positioned and oriented inside the room. We achieve this by observing that the echoes correspond to virtual or image sources, that we can exploit after correctly assigning them to walls.

It is somewhat surprising to note that the echoes help in the calibration despite not knowing where they are coming from. Supposing that the positions of the walls are unknown, the location of the source is unknown, and the locations of all the microphones are unknown, we are still able to estimate all these parameters. Particularly useful byproducts of our algorithm are partial or complete information about the room shape (as we also localize virtual sources), and the array's absolute position in the room, not available with other calibration methods. The proposed procedure is in a way a *total calibration*—we learn about microphones, sources and reflectors. We show through numerical simulations that the algorithm can indeed localize ten microphones with a single sound source.

This work was supported by an ERC Advanced Grant – Support for Frontier Research – SPARSAM Nr: 247006.

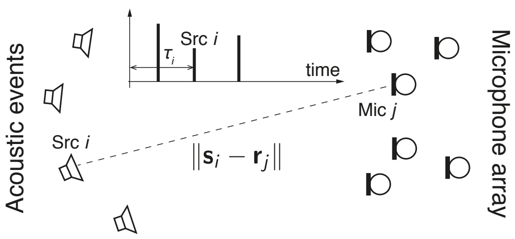

Fig. 1: Calibration without echoes.

2. USING ECHOES FOR CALIBRATION

2.1. Anechoic Calibration

As a building block in our approach, we use an algorithm for anechoic calibration. Any of the algorithms mentioned in the introduction will do. Consider the situation as in Fig. 1, and assume that the sources produce some impulsive sound that the microphones record, and whose time-of-arrival (TOA) we can estimate (up to a possibly unknown offset). Assume further that the microphones are synchronized.

We denote the source positions by $\{\mathbf{s}_k\}_{k=1}^K$, and the microphone positions by $\{\mathbf{r}_m\}_{m=1}^M$. An offset time τ_k is associated with the kth source. Then we can express the measurements as

$$\vartheta_{km} = c\,\tau_k + \|\mathbf{s}_k - \mathbf{r}_m\|_2. \tag{1}$$

We collect the measurements in a matrix $\mathbf{\Theta} = \left[\vartheta_{km}\right]_{k,m=1}^{K,M}$.

As announced, we assume the existence of a module, a black box as far as we are concerned, that we denote **Calibrate**, and that computes the estimates of the unknown microphone and source locations $\mathbf{R} \overset{\text{def}}{=} [\mathbf{r}_1, \ \ldots, \ \mathbf{r}_M], \mathbf{S} \overset{\text{def}}{=} [\mathbf{s}_1, \ \ldots, \ \mathbf{s}_K]$, and offsets $\boldsymbol{\tau} = [\tau_1, \ \ldots, \ \tau_K]^T$ from $\mathbf{\Theta}$. We can write

$$(\widehat{\mathbf{R}}, \widehat{\mathbf{S}}, \widehat{\boldsymbol{\tau}}, \varepsilon) = \textsf{Calibrate}(\mathbf{\Theta}), \tag{2}$$

where $\varepsilon \geq 0$ denotes some measure of fit. The measure of fit is computed as the discrepancy between the measured data and the data that would have been generated by sensors at estimated positions,

$$\varepsilon = \sum_{k=1}^{K}\sum_{m=1}^{M} \left|\vartheta_{km} - (\|\widehat{\mathbf{s}}_k - \widehat{\mathbf{r}}_m\|_2 + c\,\widehat{\tau}_k)\right|^2 \tag{3}$$

If $\widehat{\mathbf{R}}$, $\widehat{\mathbf{S}}$ and $\widehat{\boldsymbol{\tau}}$ perfectly generate $\mathbf{\Theta}$, then $\varepsilon = 0$.

Any algorithm behind the **Calibrate** component is associated with a certain minimal number of microphones and sources required for estimation, call them $M_{\min}$ and $K_{\min}$. Often, $K_{\min}$ is a (non-increasing) function of $M_{\min}$. A particular consequence of this is that the minimal number of columns of $\mathbf{\Theta}$ is $M_{\min}$, and the smallest corresponding number of rows $K_{\min}$. When the source offsets are known or all equal, we can swap $M_{\min}$ and $K_{\min}$ by invoking the duality.

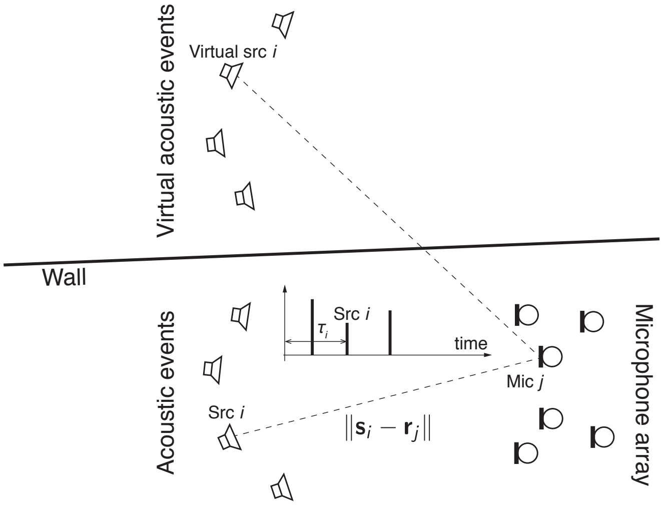

Fig. 2: Calibration with echoes.

2.2. Indoor Calibration

We observe that in a room, or more generally, in the presence of acoustic reflectors, the sources $\{\mathbf{s}_k\}$ generate reflections, and the reflections are equivalent to virtual sources (mirror images of the real sources across corresponding walls). According to the image source model [9], [10], a source at position s generates first-order virtual sources at positions

$$\mathbf{im}_i(\mathbf{s}) = \mathbf{s} + 2\langle \mathbf{p}_i - \mathbf{s}, \mathbf{n}_i\rangle \mathbf{n}_i, \tag{4}$$

where i indexes the wall, $\mathbf{n}_i$ is the outward normal associated with the ith wall, and $\mathbf{p}_i$ is any point belonging to the ith wall. Analogously, we compute image sources corresponding to higher order reflections,

$$\mathbf{im}_j(\mathbf{im}_i(\mathbf{s})) = \mathbf{im}_i(\mathbf{s}) + 2\langle \mathbf{p}_j - \mathbf{im}_i(\mathbf{s}), \mathbf{n}_j\rangle \mathbf{n}_j. \tag{5}$$

This means that we get additional sources for free. Normally, we would consider the echoes to be an annoyance, because we do not know where the virtual sources are located (the real source position is unknown and the room is unknown). But note that in the blind calibration scenario, we do not know the locations of the real sources either. Thus, virtual sources at unknown locations are just as good as real sources at unknown location. The phenomenon is illustrated in Fig. 2 for a single wall.

A challenge that appears in this setting is that we cannot address each virtual source individually—they are not labeled, and with multiple walls they are heard by microphones in different orders. This problem does not appear with real calibration events, as they are well separated in time. We need to label the echoes by performing *echo sorting*, introduced in [11]. There, the need to disambiguate echoes (virtual sources) arises when estimating the shape of a room from sound. However, the problems are not the same—in the scenario therein, the authors assume they know the relative geometry of the microphone array. In the calibration problem, we do not know it. This means that the minimal number of microphones and the minimal number of sources will be higher, as reflected by $K_{\min}$ and $M_{\min}$.

The principle at play is similar to the one used in [11]: Provided we have enough noiseless measurements ϑ_{km}, the equations (1) yield a unique solution for locations and offsets. That is, these are the only $\widehat{\mathbf{R}}, \widehat{\mathbf{S}}$ and $\widehat{\boldsymbol{\tau}}$ that could have generated $\boldsymbol{\Theta}$. Depending on the solution strategy (e.g. solving an optimization program), any wrong permutation or assignment of echoes will lead to an unsolvable system (1), or will yield a wrong solution that cannot recreate the measurements $\boldsymbol{\Theta}$.

The goal is to find the best fit among all possible echo assignments. This can be achieved by running **Calibrate** for different echo assignments, and taking as the correct assignment the one with the smallest ε. The described procedure is summarized in Algorithm 1.

Performing the combinatorial search is feasible for small array sizes. For large arrays, however, the number of combinations becomes too big, and we need to do something else. In this case, we can bootstrap the method by first running it for one or more sub-arrays of the big array. Depending on the target application, we might even have an idea about groups of microphones that are spatially close (this will be the case for large fixed arrays). Knowing which microphones are close in space is relevant, as proximity makes it more likely that the microphones will have picked up the same echoes. In spatially large arrays, it is not guaranteed that all the microphones will hear all the echoes.

The estimation can be performed one acoustic event at a time (e.g. a finger snap). This is useful, as we know that the time offset τ will be the same for all virtual events corresponding to a single real event (and it will be equal to the offset of that event). Structured information like this can be exploited in the design of the **Calibrate** module.

2.3. Minimal Infrastructure for Calibration

We can use a degree-of-freedom counting argument as in [2] to determine the smallest number of microphones and sources necessary for the calibration. Every microphone brings in three unknowns (x, y and z coordinates), while every source brings in four unknowns (coordinates and the offset τ). On the other hand, every source gives us M TOA measurements. The total number of measurements is then MK, and we need this number to be larger than the total number of unknowns, $3M+4K$. Note further that we can fix the location of one microphone and the rotation of the remaining points around this microphone. This takes out a total of six degrees of freedom, resulting finally in

$$K \geq \left\lceil \frac{3M-6}{M-4} \right\rceil . \tag{6}$$

The only remaining ambiguity is the 1-bit reflection ambiguity. The relationship between the minimal number of microphones and sources is always a property of the method used to calibrate.

Algorithm 1 BASIC INDOOR CALIBRATION

```
Input:    ▷ Microphone recorded signals ${y_m(t)}_{m=1}^{M}$
Output:   ▷ Estimated microphone positions $\widehat{\mathbf{R}}$
          ▷ Estimated source positions $\widehat{\mathbf{S}}$
          ▷ Estimated source offsets $\widehat{\boldsymbol{\tau}}$
Peak picking:
    ▷ For every $y_m(t)$ find the set of peaks (echoes), $T_m$
Initialization:
    ▷ $\varepsilon_{\text{best}} \leftarrow \infty$
For every feasible echo assignment across $\{T_m\}$:
    ▷ Create the corresponding matrix $\boldsymbol{\Theta}_i$
    ▷ $(\widehat{\mathbf{R}}, \widehat{\mathbf{S}}, \widehat{\boldsymbol{\tau}}, \varepsilon) = \textsf{Calibrate}(\boldsymbol{\Theta}_i)$
    ▷ If $(\varepsilon < \varepsilon_{\text{best}})$, then $(\mathbf{R}_{\text{best}}, \mathbf{S}_{\text{best}}, \boldsymbol{\tau}_{\text{best}}, \varepsilon_{\text{best}}) \leftarrow (\widehat{\mathbf{R}}, \widehat{\mathbf{S}}, \widehat{\boldsymbol{\tau}}, \varepsilon)$
End For
```

We use this example to show that something remarkable can happen in a room. Suppose that $M = 10$. In this case, we compute that $K \geq 4$—we need at least four sources. Now imagine that in a room we have a single acoustic event, and that we can get at least three echoes. Together with the direct sound, we get enough measurements (real and virtual) to actually calibrate the microphone array, and to determine its absolute orientation with respect to the walls. This is true in spite of the fact that we do not know the room, the microphone locations, the source location nor the source timing. In this case we only need to estimate a single emission time, as it will be the same for all the image sources.

3. PRACTICAL ALGORITHM

3.1. Reducing the number of combinations

The main drawback of the proposed algorithm is the combinatorial search. Especially in the case of large arrays, the number of combinations becomes too large to test them all. The following heuristics can be employed in order to reduce the number of combinations,

i) Perform the estimation for sub-arrays,

ii) Combine the echoes only within a temporal window corresponding to the array size,

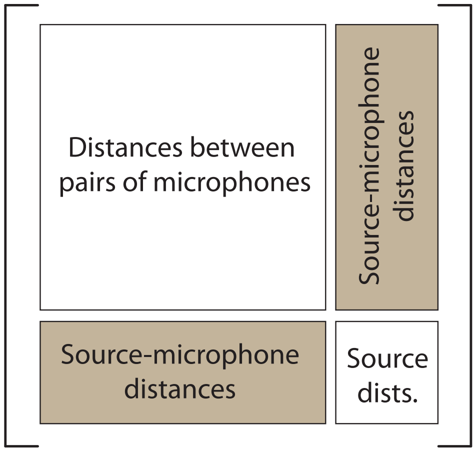
Distances between pairs of microphones
Source-microphone distances
Source-microphone distances
Source dists.

Fig. 3: Structure of the EDM used to fine-tune the estimate.

iii) Assume only a small number of echo swaps can occur per microphone,

iv) Assume echo swaps occurred at a limited number of microphones,

v) Normalize by the decay and discriminate first-order peaks by the magnitude,

vi) Order the echo assignments by the likelihood and stop as soon as we get a score below a prescribed threshold.

3.2. Dealing with Noise

For the numerical experiments in this paper, we implemented the **Calibrate** module based on [1]. A nice feature of that algorithm is that in its basic version it is fast, and it gives a good initial estimate of the unknown times and locations.

To further optimize the estimate, we propose an iterative algorithm based on Euclidean distance matrices (EDM) [12] and multi-dimensional scaling (MDS) [13].

Let $\{\mathbf{x}_i\}_{i=1}^{m+n}$ list the points corresponding to sources and to microphones, so that $\mathbf{x}_i = \mathbf{r}_i$ for $1 \leq i \leq M$ and $\mathbf{x}_i = \mathbf{s}_{i-M}$ for $(M+1) \leq i \leq (M+K)$. Denote by $\mathbf{D}$ the Euclidean distance matrix (EDM) corresponding to the point set $\{\mathbf{x}_i\}$, that is, $\mathbf{D} = (d_{ij})$, where $d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|^2$. The structure of $\mathbf{D}$ is illustrated in Fig. 3.

The main ingredient of the solution is an algorithm for MDS. MDS aims to embed a set of points in $\mathbb{R}^n$ given their noisy pairwise distances. Among the many approaches to MDS we choose to minimize the cost function called s-stress. Given a noisy EDM $\widetilde{\mathbf{D}}$, we obtain a denoised matrix as

$$\mathrm{MDS}(\widetilde{\mathbf{D}}) \overset{\text{def}}{=} \underset{\mathbf{D} \in \mathrm{EDM}^3}{\arg\min} \sum_{ij} \left(d_{ij}^2 - \widetilde{d}_{ij}^2 \right)^2 . \tag{7}$$

By EDM^3 we denote the set of EDMs generated by point sets in 3D. This cost function is not convex. Nevertheless, it was shown that a coordinate-alternating approach to minimization almost always achieves the global optimum [14].

Algorithm 2 EDM-BASED UNFOLDING

```
Input:   ▷ Initial source-microphone distances
         ▷ Initial inter-microphone distances
Output:  ▷ Estimates of $\{\widehat{\mathbf{s}}_k\}_{k=1}^{K}$ and $\{\widehat{\mathbf{r}}_m\}_{i=m}^{M}$
Initialization:
    ▷ Construct the matrix $\mathbf{D}$ from the input data (source-microphone distances into the shaded part, and inter-microphone, and inter-source distances in the non-shaded parts)
Repeat
    ▷ Set the elements of $\mathbf{D}$ corresponding to source-mic distances to input values (shaded regions in Fig. 3)
    ▷ Find the closest EDM to $\mathbf{D}$, $\mathbf{D} \leftarrow \arg\min s(\mathbf{D})$
Until convergence
    ▷ Estimate the positions of points generating $\mathbf{D}$ as $\{\mathbf{x}_i\}_{i=1}^{M+K} = \underset{\{\mathbf{x}_i \in \mathbb{R}^3\}}{\arg\min} \sum_{ij} \left( \|\mathbf{x}_i - \mathbf{x}_j\|^2 - \widetilde{d}_{ij}^2 \right)^2$, and extract $\widehat{\mathbf{R}}$ and $\widehat{\mathbf{S}}$
```

We assume that the source-microphone distances were more accurately estimated than the inter-microphone distances. Thus the iteration of the proposed algorithm consists in enforcing the elements of the matrix corresponding to source-microphone distances, and then *relaxing* the matrix using MDS. We empirically observe that the described procedure (Algorithm 2) considerably improves the initial estimate.

4. NUMERICAL EXPERIMENT—SINGLE SOURCE

We ran numerical simulations using only one acoustic event to localize ten microphones. In this case, all of the sources (real or virtual) have the same offset τ.

We simulated a shoebox room with dimensions $W = 5$ m, $L = 6$ m, $H = 3$ m, using the image source model, up to second-order reflections. We experimented with random microphone array geometries and with different numbers of microphones. The algorithm used for echo sorting and for the final reconstruction is the combination described in Section 3. Room acoustics were simulated using the image source model up to second-order reflections, and the first six echoes were used for estimation.

Simulations confirm that it is possible to obtain accurate estimates of microphone positions by using only a single source. The room shape or dimensions are considered unknown by the algorithm. Despite this, we obtain a full reconstruction of the source and microphone locations, as well as their absolute pose inside the room (more precisely, relative to the localized image sources). We also obtain the positions of walls corresponding to these image sources. Two reconstruction examples are shown in Fig. 4 (A) and (B), for random microphone arrays comprising ten microphones. Fig.

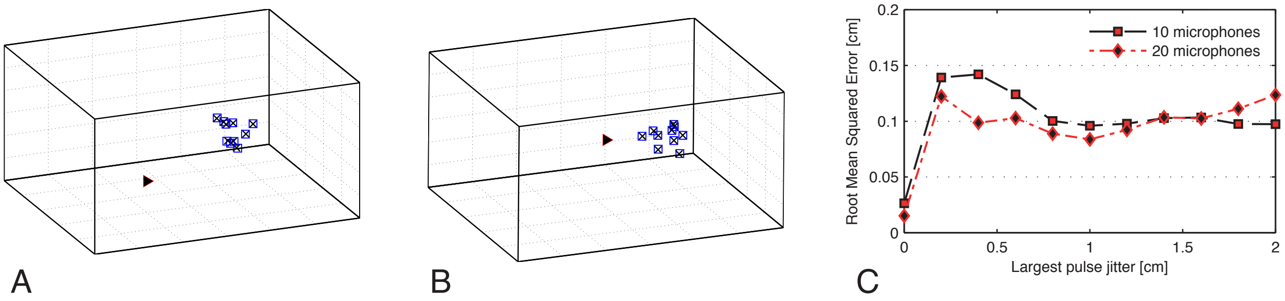

Fig. 4: (A) and (B) Two typical reconstruction results with $M = 10$ microphones randomly placed inside a box approximately $1\text{m} \times 1\text{m} \times 0.5\text{m}$ large. We emphasize that the room dimensions (5m × 6m × 3m) and the room shape is not assumed known. Red-black triangle represents the source location. Small black crosses are true microphone locations, while blue squares denote the estimated locations. (C) Accuracy of microphone localization with jittered pulses. The jitter that was added to pulses was drawn from $\mathcal{U}[-d, d]$, where d is indicated on the abscissa in [cm]. The room was of the same dimensions as in (A) and (B).

4 (C) shows the root-mean-squared error for the estimates of microphone positions against the amount of jitter. It can be seen that the algorithm performs stably in moderate jitter.

5. CONCLUSION

We presented the proof-of-concept of constructive use of echoes for microphone array calibration. Interpreting echoes as virtual sources allows us to reduce the number of sources required to calibrate the array. In the extreme case, it is possible to calibrate using only a single acoustic event such as a finger snap, even without knowing the room. To the best of our knowledge, this is the first description of such a possibility.

The main line of future work concerns reducing the number of echo assignments to test. Furthermore, we intend to design **Calibrate** modules adapted for the specific case of equal time offsets.

References

[1] M. Pollefeys and D. Nister, "Direct computation of sound and microphone locations from time-difference-of-arrival data," in *Proc. IEEE Intl. Conf. on Acoust., Speech and Signal Process.*, Las Vegas, 2008, pp. 2445–2448, IEEE.

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

[3] E. Weinstein, K. Steele, A. Agarwal, and J. Glass, "LOUD: A 1020-node microphone array and acoustic beamformer," in *International Congress on Sound and Vibration*, 2007.

[4] J. M. Sachar, H. F. Silverman, and W. R. Patterson, "Microphone position and gain calibration for a large-aperture microphone array," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 42–52, Jan. 2005.

[5] V. C. Raykar and R. Duraiswami, "Automatic position calibration of multiple microphones," in *Proc. IEEE Intl. Conf. on Acoust., Signal and Speech Process.*, Montreal, 2004, pp. 69–72, IEEE.

[6] M. Crocco, A. D. Bue, and V. Murino, "A Bilinear Approach to the Position Self-Calibration of Multiple Sensors," *IEEE Trans. Signal Process.*, vol. 60, no. 2, pp. 660–673, 2012.

[7] P. H. Schönemann, "On metric multidimensional unfolding," *Psychometrika*, vol. 35, no. 3, pp. 349–366, 1970.

[8] S. Thrun, "Affine Structure From Sound," in *Proc. Conf. Neural Inf. Process. Sys. (NIPS)*, Cambridge, MA, 2005, MIT Press.

[9] 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, 1979.

[10] J. Borish, "Extension of the Image Model To Arbitrary Polyhedra," *J. Acoust. Soc. Am.*, vol. 75, no. 6, pp. 1827–1836, 1984.

[11] I. Dokmanić, R. Parhizkar, A. Walther, Y. M. Lu, and M. Vetterli, "Acoustic echoes reveal room shape," *Proc. Natl. Acad. Sci.*, vol. 110, no. 30, June 2013.

[12] J. Dattorro, *Convex Optimization & Euclidean Distance Geometry*, Meboo, 2011.

[13] W. S. Torgerson, "Multidimensional scaling: I. Theory and method," *Psychometrika*, vol. 17, no. 4, pp. 401–419, Dec. 1952.

[14] R. Parhizkar, *Euclidean distance matrices: Properties, algorithms and applications*, Ph.D. thesis, EPFL, Lausanne, 2013.