**Circuits and Systems**
Mekelweg 4,
2628 CD Delft
The Netherlands
http://ens.ewi.tudelft.nl/

CAS-2018-00

# M.Sc. Thesis

## Room Geometry Estimation from Acoustic Echoes

**Raissa Lynn B.Sc.**

### Abstract

Estimating a room geometry using multiple microphones rises an echoes labeling problem. Two recent methods called the graph-based and the subspace-greedy methods have shown their capability in solving this problem. The graph-based method attains a good accuracy but suffers in maintaining the computational cost when the number of microphones is larger than 7. On the other hand, the subspace-greedy method provides suboptimal accuracy with much lower computational time. Here we construct the hybrid combination methods using those two baseline methods by interchanging their intermediate steps: the refinement step and the source localization step. To assess their practicability in a real-life application such as virtual reality games and robot navigation, the performance of these hybrid methods were tested against the close microphones arrangement on the sphere's surface. However, this new microphones' constellation brings up a low dimensional problem. To deal with this matter, we use the weighted least squares as the source localization procedure. Finally, experiments on synthetic squared distance data demonstrate the feasibility of all hybrid methods for estimating the room geometry with centimeter precision within seconds.

TU Delft

Faculty of Electrical Engineering, Mathematics and Computer Science

Delft University of Technology

# Room Geometry Estimation from Acoustic Echoes

## My Subtitle

Thesis

submitted in partial fulfillment of the
requirements for the degree of

Master of Science

in

Electrical Engineering

by

Raissa Lynn B.Sc.
born in Jakarta, Indonesia

This work was performed in:

Circuits and Systems Group
Department of Microelectronics & Computer Engineering
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology

TU Delft

Delft University of Technology

Delft University of Technology
Department of
Microelectronics & Computer Engineering

The undersigned hereby certify that they have read and recommend to the Faculty of Electrical Engineering, Mathematics and Computer Science for acceptance a thesis entitled **"Room Geometry Estimation from Acoustic Echoes"** by **Raissa Lynn B.Sc.** in partial fulfillment of the requirements for the degree of **Master of Science**.

Dated: 20 September 2018

Chairman: ______________________
dr.ir. R. Heusdens

Advisor: ______________________
dr.ir. R. Heusdens

Committee Members: ______________________
dr. O.E. Scharenborgr

______________________
dr. N.D. Gaubitch

______________________
dr. J.A. Martinez Castaneda

# Abstract

Estimating a room geometry using multiple microphones rises an echoes labeling problem. Two recent methods called the graph-based and the subspace-greedy methods have shown their capability in solving this problem. The graph-based method attains a good accuracy but suffers in maintaining the computational cost when the number of microphones is larger than 7. On the other hand, the subspace-greedy method provides suboptimal accuracy with much lower computational time. Here we construct the hybrid combination methods using those two baseline methods by interchanging their intermediate steps: the refinement step and the source localization step. To assess their practicability in a real-life application such as virtual reality games and robot navigation, the performance of these hybrid methods were tested against the close microphones arrangement on the sphere's surface. However, this new microphones' constellation brings up a low dimensional problem. To deal with this matter, we use the weighted least squares as the source localization procedure. Finally, experiments on synthetic squared distance data demonstrate the feasibility of all hybrid methods for estimating the room geometry with centimeter precision within seconds.

# Acknowledgments

First, I would like to thank Jesus for His grace and faithfulness that make me who I am today. Many thanks to my supervisor Dr. R. Heusdens for his guidance, insight, and powerful feedback throughout my thesis. It is such an honor for me to do my thesis with you. To all lecturers who I met in TU Delft, thank you for giving me unforgettable experiences and great knowledge. Mario, thank you for your time in helping me to understand this topic and giving me some good advice. Finally, I would like to express my gratitude to my dad, Gwan Hok, my mom, Trivena, and my sister, Hilary for their unending support and trust towards me.

Raissa Lynn B.Sc.
Delft, The Netherlands
20 September 2018

# Contents

# List of Figures

# List of Tables

# Introduction 1

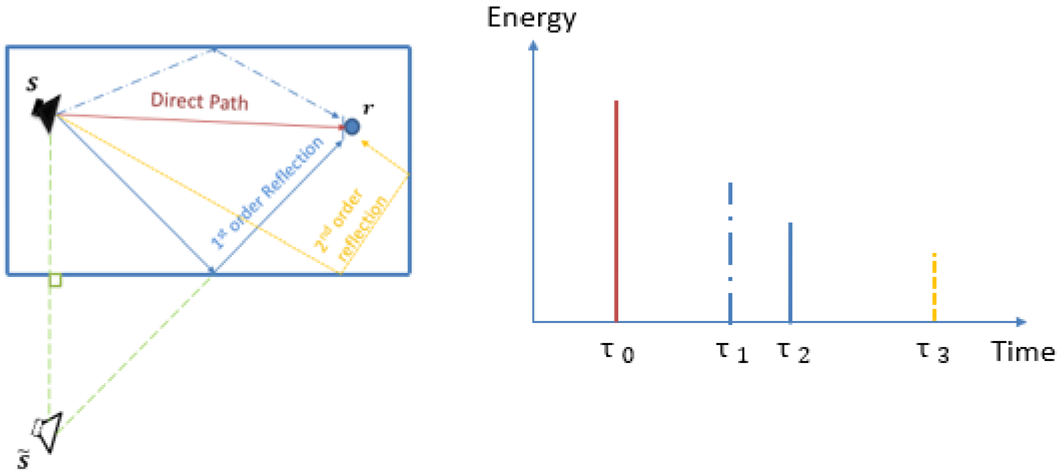


Figure 1.1: Left: A two dimensional top view of a room with a microphone ($r$) and a source ($S$). Right: An ideal room impulse response (RIR) with peaks correspond to the sound paths in the room. Each $\tau$ in $x$ axes defines the time of arrival (TOA) of each path on the microphone

Advance development of reality technologies changes the way we perceive our environment. For example, using augmented reality technology people can see virtual objects on top of real ones. Even in the latest mixed reality technology, users can interact with the virtual objects in a real scene. This condition means that the degree of immersion between a human and the reality technology should be high and dynamic. Thus, a combination of visual and auditory information is required since the visual perception is significantly augmented with the matched sound stimuli. Moreover, a good room estimation result can benefit several applications such as scene reconstruction, spatial sound rendering [2], and objects or sources localization [3]. This fact pushes the room geometry estimation to play an important role in improving the reality technology.

Consider a two dimensional top view of a room as shown in the left part of Fig.1.1. The first arrived signal in the microphone is called a direct path. Signals that come after this are called echoes or reflections. All the sound paths in the room are also indicated as peaks in the room impulse response (The right figure in Fig. 1.1). It shows that the received signal energy decreases as the travel distance increases. As we can see from Fig.1.1, the locations of the reflections contain spatial cues of the enclosures. Adopting the image source method to model the sound propagation inside a room, we can replace a reflected path with a direct path from an image source outside the enclosure to the microphone [4]. This situation is illustrated in Fig.1.1 where $S$

and $\tilde{S}$ denotes the real source and image source respectively. If the location of the reflections, i.e., image sources and real sources are known, the normals and the location of the reflective surfaces (walls) can be deducted by a simple geometric relation. Thus, in order to infer the room geometry, our main task is to find the location of the image sources.

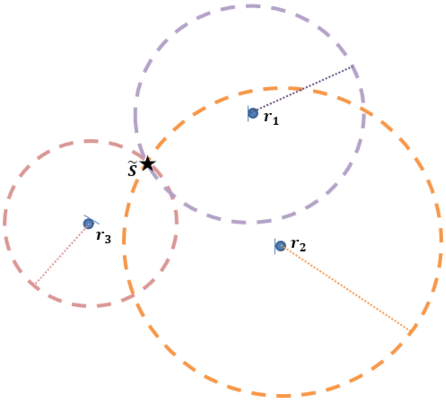


Figure 1.2: Source localization from intersection of three distances

In two dimensions case when a microphone receives the direct path contribution, we know that the source lies on a sphere with radius related to the time of arrival. To resolve the ambiguity, we need at least three independent observations (microphones). Figure 1.2 illustrates that the image source location ($\tilde{s}$) is the intersection of three circles with radii equal to the distance between each microphone and the image source. For a three dimensional case, the minimum number of microphones that is required is four. However, adding more microphones does not guarantee that the position of image sources can be found since echoes (distances) for locating them are ambiguous. In the RIR representation, we do not know which peaks match the corresponding image source. This ambiguity issue is also known as an echo labeling problem and is depicted in Fig. 1.3. In this picture, the order of echoes that arrive at each microphone is swapped so a simple assumption that the first echo in the RIR of both microphones belongs to the green wall will lead to a false room geometry estimation. Therefore, a correct assignment between echoes and its responsible walls is necessary.

There are some available methods for solving the echo labeling problem. In [5], Dokmanic et al. exploited the properties of Euclidean distance matrices which leads to a multidimensional scaling (MDS) method for correctly assigned the wall using one real source and five microphones. Though this method gives accurate results, the computational cost is very high which makes it is not suitable for real-time application. Alternatively, Jager et al. in [6] formulated the echo labeling problem as a graph problem which can be solved in a modest time with the small number of microphones ($M < 7$) while retaining the same accuracy as [5]. The latest approach based on a subspace technique and a greedy echo selection procedure by Coutino et al [7] further reduced the computational complexity but gave a sub-optimal performance. Consequently, a tradeoff between computational cost and the accuracy is an inevitable issue. In both Jager's and Coutino's methods, the microphones for performing the measurements are

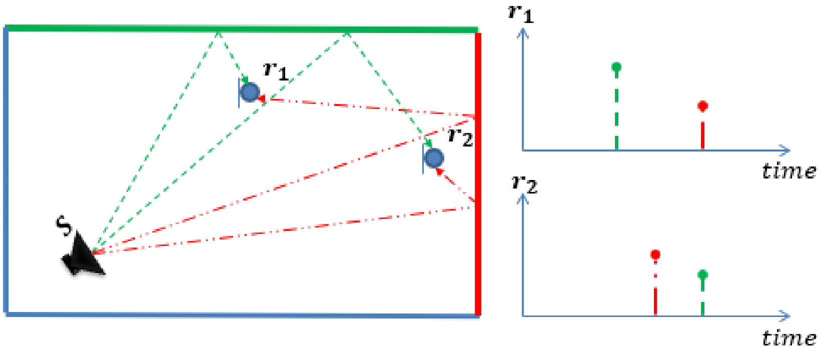


Figure 1.3: Echo labeling problem

randomly distributed inside a room.

For an application of room geometry estimation in the area such as virtual reality and robot navigation, the environmental awareness is a crucial part. Imagine when the virtual reality user or the robot moves from one room to another room. In this case, the room geometry estimation must be done in a continuous, real-time, and adaptable manner to achieve a good performance. These requirements are tantamount to having mobile sensors with reasonable size that can be carried by the user (a person or a robot). For example, in a virtual reality game, the user can be equipped with a visual glass and helmet like sensors (microphones) while in the robot navigation, the microphones can be located on the head of the robot.

## 1.1 Research Statement and Outline

In this thesis, the following general research question is addressed:

> *What is the most efficient and accurate technique for estimating the room geometry based on the graph and subspace method for the specific case of close microphones arrangement?*
> *with assumptions that the RIR of a shoe box room is known and there is no occlusion inside the room.*

The rest of the thesis is organized as follows: Chapter 2 describes some background theories. Chapter 3 introduces the methods by Jager and Coutino in more detail, also provides its evaluation result. In Chapter 4, the implementation of both methods for helmet microphones configuration is presented. Finally, Chapter 5 closes this thesis with the conclusion and suggestion for future work.

# Background Theory 2

This chapter outlines supporting theories of some basic tools for estimating the room geometry.

## 2.1 Room Geometry Estimation Pipeline

The room geometry estimation problem can be divided into different subproblems as shown in Figure 2.1.

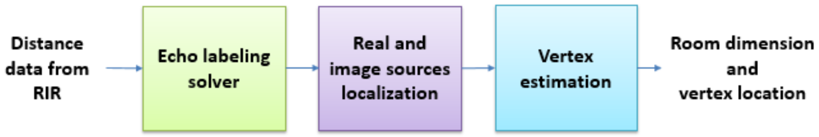

Distance data from RIR
Echo labeling solver
Real and image sources localization
Vertex estimation
Room dimension and vertex location

Figure 2.1: Room geometry estimation pipeline

The input of the room geometry estimation pipeline is the distance data obtained from the room impulse response. In the initial step, we solve the echo labeling problem. The output of this step is the true labeled distance data. In the next step, this data is processed to localize the real and image source positions. Finally, based on the knowledge of the source positions and the microphone positions, the room boundaries can be revealed.

The next section will deal with some basic theories that serve as a guide for the reader to understand this topic clearly. Moreover, some related works in solving the room geometry estimation problem are presented in the last section of this chapter.

## 2.2 Background Theory

### 2.2.1 Room Impulse Response (RIR)

A room (enclosure) can be considered as an acoustic system which transforms any sound signal that fed into it. The acoustic response which caught by receivers (microphones) in a room varies according to the position of the sources, receivers, and acoustical condition of room surfaces. According to the concept of geometrical room acoustics (as described in [8]), the transformation that is experienced by the sound signal is the result of multiple time reflections at the room boundaries, so it is also related with how the sound signal propagates in a room. Thus, a received signal at the microphone would be the superposition of many replicas of the original signal. Each of them has its particular strength and is delayed by its particular traveling time. This phenomenon

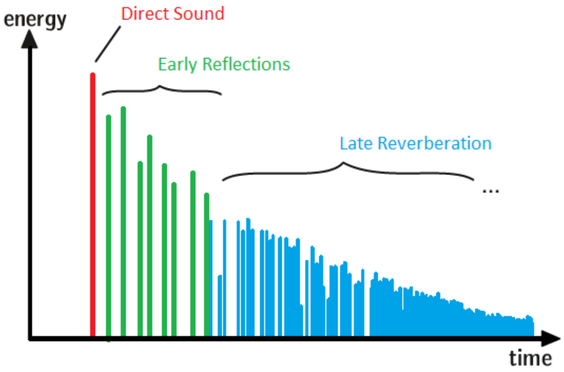


Figure 2.2: Energy vs Time Curve of Room Impulse Response

can be well understood using the response of a room to an ideal impulse signal, i.e., room impulse response (RIR) which can be expressed as

$$h(t) = \sum_{n} A_n \delta(t - \tau_n). \tag{2.1}$$

Ideally $h(t)$ is a train of delta pulses, each pulse corresponds to an echo. $A_n$ and $\tau_n$ are the signal strength and delay time of the $n^{th}$ echo respectively.

Any sound emitted in a room will be affected by the RIR. The resulting output signal $g(t)$ is the convolution of the emitted sound signal $s(t)$ and the RIR $h(t)$. In equation form, it can be expressed as

$$g(t) = \int_{-\infty}^{\infty} s(\tau)h(t - \tau)d\tau = s(t) * h(t). \tag{2.2}$$

A simple illustration of an RIR is depicted in Fig 2.2. Typically, the RIR is divided into three major parts :

- Direct sound: The first arriving sound at the microphone.
- Early reflection: Discrete, sparse reflective sound signal from nearby surfaces.
- Late reverberation: Densely populated reflective sound signal.

From the RIR we can extract the time delay of arrival (TOA) related to the direct sound and early reflections (echoes). These TOAs can be translated into distances between the sources and the microphones through

$$\tau_n = \frac{d_n}{c} \quad n = 0,1,2,\cdots, \mathbf{n} \ , \tag{2.3}$$

where $\tau_n$ is the TOA, $c$ is the sound velocity, $d_n$ is the distance, index $n = 0$ corresponds to the direct sound while $n = 1, 2, \cdots, \mathbf{n}$ correspond to the echoes. We can link this information with the image source model that will be described in the next section to locate the room boundaries.

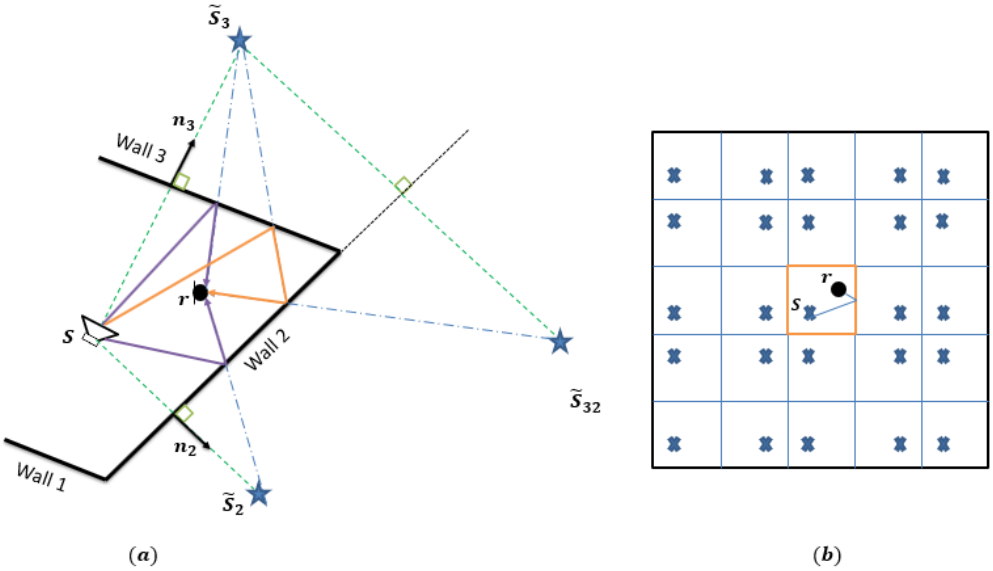


Figure 2.3: (a) Image source model for the first and second order reflections. Vector $\boldsymbol{n}_i$ is the unit normal associated with the $i^{th}$ wall. $\tilde{\boldsymbol{s}}_i$ denotes the image sources w.r.t the $i^{th}$ wall. $\tilde{\boldsymbol{s}}_{ij}$ is the image source for the second order echo. (b) Repeated image source pattern for a box shaped room (the orange box)

### 2.2.2 Image Source Models

The image source model is categorized as a ray-based method which based on geometrical room acoustics. In geometrical room acoustics, the concept of sound waves is substituted by the concept of sound rays [8]. Provided a homogeneous medium, the sound ray travels along a straight line with constant energy. If a sound ray strikes a solid surface, it will be reflected. The reflection law is the same as from optics. Any typical wave effects such as diffraction and interference are neglected in geometrical room acoustics. Then, it is evident that geometrical room acoustics can only reflect a partial aspect of the acoustical phenomena inside a room. However, this disadvantage is covered by its conceptual simplicity and practical computation.

The concept of the image source model is straightforward: for each reflective surface, a virtual sound source is produced by mirroring the sound source across the corresponding surface (wall) as illustrated in Fig. 2.3(a). Therefore, the echo can be modeled by the direct sound emanating from the virtual source. If a reflected sound ray strikes the second wall, the continuation of the sound path can be found by repeating the mirroring process. The image sources method is very efficient for a box-shaped environment due to the rectilinear symmetries of a box [4]. Figure 2.3(b) shows this situation.

In a mathematical form, the position of an image source can be represented by

$$\tilde{\boldsymbol{s}}_i = \boldsymbol{s} - 2\langle \boldsymbol{s} - \boldsymbol{p}, \boldsymbol{n}_i \rangle \boldsymbol{n}_i \ , \tag{2.4}$$

where $\boldsymbol{s}$ denotes the source position , $\tilde{\boldsymbol{s}}_i$ is the position of the image source correspond to the $i^{th}$ wall, $\boldsymbol{n}_i$ is the wall unit normal vector and $\boldsymbol{p}$ is an arbitrary point on the wall. See Fig.2.4 for an illustration.

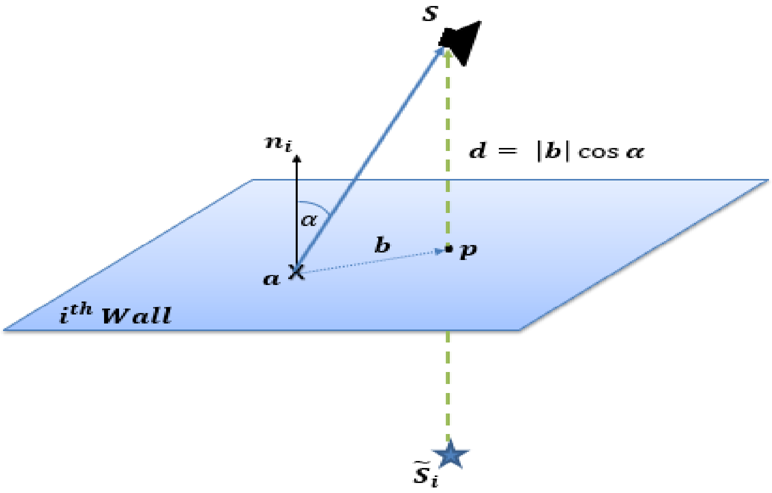


Figure 2.4: An image source model generation

### 2.2.3 Room Reconstruction

To localize the wall , we employ the loudspeaker-image bisection (LIB) algorithm [9]. The $i^{th}$ wall ($\boldsymbol{W}_i$) is a plane that bisecting the line from the source $\boldsymbol{s}$ to its image source $\tilde{\boldsymbol{s}}_i$ (the dotted green line in Fig. 2.4). Their midpoint $\boldsymbol{p}$ lies on the plane. Once both the true source ($\boldsymbol{s}$) and the first order image source position ($\tilde{\boldsymbol{s}}_i$) are available, the unit vector normal to the $\boldsymbol{W}_i$ can be computed:

$$\boldsymbol{n}_i = \frac{\boldsymbol{s} - \tilde{\boldsymbol{s}}_i}{\|\boldsymbol{s} - \tilde{\boldsymbol{s}}_i\|} \quad , \tag{2.5}$$

where $\| \cdot \|$ represents the Euclidean norm. The midpoint $\boldsymbol{p}$ is defined as

$$\boldsymbol{p} = \frac{\boldsymbol{s} + \tilde{\boldsymbol{s}}_i}{2} \quad . \tag{2.6}$$

Hence, using this midpoint and the normal vector $\boldsymbol{n}_i$, the plane $\boldsymbol{W}_i$ can be defined in homogeneous coordinates as:

$$\boldsymbol{W}_i = \left[\boldsymbol{n}_i^T, -\boldsymbol{p}^T \boldsymbol{n}_i\right]^T \ . \tag{2.7}$$

Any points $\boldsymbol{a}$ that lies on the $\boldsymbol{W}_i$ must satisfy:

$$\langle \boldsymbol{n}_i, \boldsymbol{a} - \boldsymbol{p} \rangle = 0 \, . \tag{2.8}$$

The vertices of the room can be found at the intersections of the boundary planes. The accuracy of the estimated $i^{th}$ wall position improves as more points on the boundary plane are involved. These points can be obtained by either moving the source or providing more sources inside the room.

### 2.2.4 Euclidean Distance Matrix (EDM)

An EDM consists of squared Euclidean distances between a set of $M$ points in an $N$-dimensional Euclidean space. Consider a matrix $\mathbf{X} \in \mathbb{R}^{N\times M}$, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_M]$ which columns represent points in $N$ dimensional Euclidean space. The squared distance between point $\mathbf{x}_i$ and $\mathbf{x}_j$ is defined by

$$d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|^2 . \tag{2.9}$$

Expanding Eq. 2.9 yields

$$d_{ij} = (\mathbf{x}_i - \mathbf{x}_j)^T(\mathbf{x}_i - \mathbf{x}_j) = \mathbf{x}_i^T\mathbf{x}_i - 2\mathbf{x}_i^T\mathbf{x}_j + \mathbf{x}_j^T\mathbf{x}_j. \tag{2.10}$$

If we calculate $d_{ij}$ for every paired point in the point set $\mathbf{X}$, we can construct a Euclidean distance matrix, $\mathit{EDM}(\mathbf{X})$ which equal to a matrix $\mathbf{D}$ with element $d_{ij}$:

$$\mathit{EDM}(\mathbf{X}) \triangleq \mathbf{D} = \mathbf{1}\mathrm{diag}(\mathbf{X}^T\mathbf{X})^T - 2(\mathbf{X}^T\mathbf{X})^T + \mathrm{diag}(\mathbf{X}^T\mathbf{X})\mathbf{1}^T, \tag{2.11}$$

diag($\mathbf{A}$) is a column vector of the diagonal entries of $\mathbf{A}$ and $\mathbf{1}$ is the all ones column vector.

Equation 2.11 reflects a very important property of the EDM that we will encounter and exploit later in this thesis to solve the echo labeling problem. In fact, $\mathbf{D}$ is a function of $\mathbf{X}^T\mathbf{X}$. Since the rank of $\mathbf{X}$ is at most $N$, so does the rank of $\mathbf{X}^T\mathbf{X}$. The remaining parts in Eq. 2.11 have rank one. By rank inequalities, the rank of the sum of matrices should less than or equal the sum of the ranks of the summands. The plausible conclusion of this condition is stated in Theorem 1.

**Theorem 1.** *The rank of an EDM ($\boldsymbol{D}$) corresponding to points in an $N$ dimensional space is:*

$$\mathit{rank}(\boldsymbol{D}) \leqslant N + 2. \tag{2.12}$$

Theorem 1 entails a fundamental matter about the dimension of the smallest affine subspace that contains the point set, i.e., the affine dimension of the point set ($\mathbf{X}$), denoted by affdim($\mathbf{X}$). As an illustration, consider points from the point set $\mathbf{X} \in \mathbb{R}^3$ that lie randomly on a plane in $\mathbb{R}^3$, the rank of the corresponding EDM is not five but four. It means that the EDM which is generated by this point set can also be generated from another set of points $\mathbf{X}' \in \mathbb{R}^2$ maintaining the same distance as in the three dimensional case. Hence, there are infinitely many points sets able to construct a given EDM.

Theorem 1 also states that the rank of an EDM is independent of the number of points that generate it. In many applications, $N$ is three or less while $M$ can be in thousands [10].

When we work with an EDM, we translate our problem or information, i.e., our point set $\mathbf{X}$ into the distance geometry. The encoding process eliminates the information about the absolute position and orientation of $\mathbf{X}$. Intuitively, the rigid transformations do not change the distances between the fixed points in $\mathbf{X}$. This fact is easily deduced from Equation 2.11 since ($\mathbf{D}$) is actually a function of the Gram matrix $\mathbf{X}^T\mathbf{X}$.

For the sake of clarity, we can show that rotations and reflections do not alter the distances. Any rotation or reflection can be regarded as an orthogonal matrix

$\boldsymbol{Q} \in \mathbb{R}^{N\times N}$ acting on the point set $\mathbf{X} \in \mathbb{R}^{N\times M}$. Thus, the Gram matrix of the rotated point set $\mathbf{X}_r = \boldsymbol{Q}\mathbf{X}$ is

$$\mathbf{X}_r^T \mathbf{X}_r = (\boldsymbol{Q}\mathbf{X})^T(\boldsymbol{Q}\mathbf{X}) = \mathbf{X}^T \boldsymbol{Q}^T \boldsymbol{Q}\mathbf{X} = \mathbf{X}^T\mathbf{X} \,, \tag{2.13}$$

here the orthogonality of the matrix $\boldsymbol{Q}$, $\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I}$ has been used. Furthermore, consider the point set $\mathbf{X}$ is translated by a vector $\boldsymbol{b} \in \mathbb{R}^N$, i.e.,

$$\mathbf{X}_t = \mathbf{X} + \boldsymbol{b}\mathbf{1}^T. \tag{2.14}$$

Observing that $\text{diag}(\mathbf{X}_t^T\mathbf{X}_t) = \text{diag}(\mathbf{X}^T\mathbf{X}) + 2\mathbf{X}^T\boldsymbol{b} + \|b\|^2\mathbf{1}$, one can easily demonstrate that this translation leaves (2.11) unchanged which proves exhibit the invariance of EDM against the rigid transformations. In summary,

$$EDM(\boldsymbol{Q}\mathbf{X}) = EDM(\mathbf{X} + \boldsymbol{b}\mathbf{1}^T) = EDM(\mathbf{X}). \tag{2.15}$$

The corollary of this invariance is the inability to reconstruct the absolute orientation of the generating point set using only the distances. Different reconstruction procedures that recover $\mathbf{X}$ from $\mathbf{D}$ lead to distinct realizations of the point set. Each of them is differentiated by the rigid transformations.

## 2.3 Related Works

This section summarizes some prior works from other literature which have addressed the problem of room geometry estimation using RIR. Most of the methods either assume prior knowledge of the RIR or impose some conditions on the microphones and sources position.

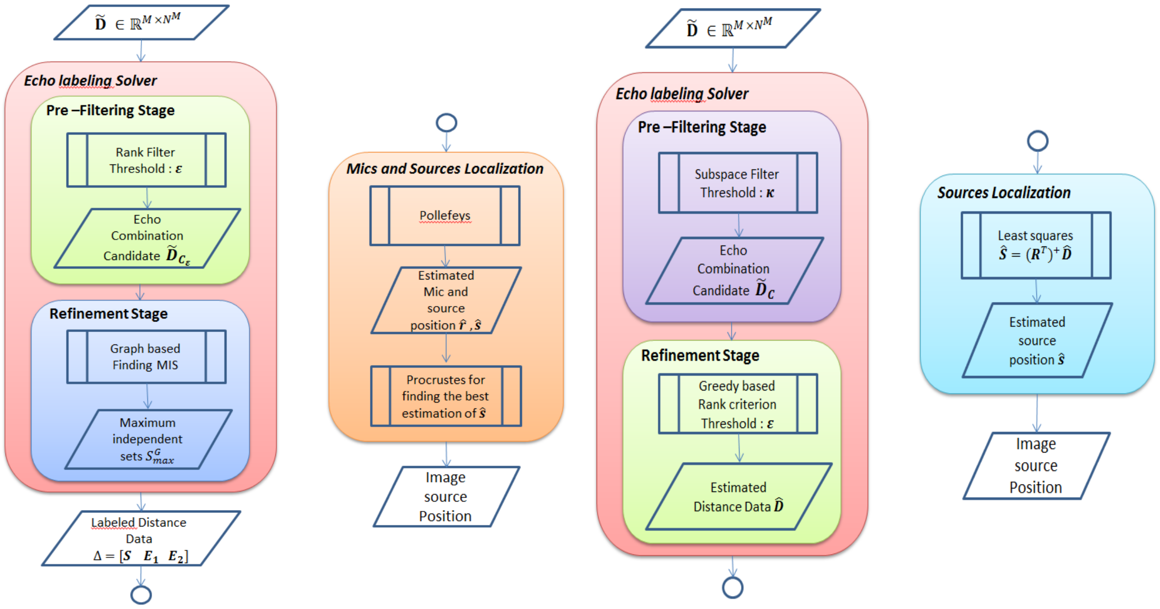
In [11], the author estimated the 2D shape of a room using a single RIR from a collocated sound source and microphone. Atonacci et al [12] localized 2D reflectors by deploying RIR from multiple microphones and a moving source, then the authors used elliptical constraints and Hough transform to finalize the result. Furthermore, [13] improved and extended the latest methods into 3D case.

In [14] the 3D room estimation problem is tackled with RIR between a small circular array of microphones and an integrated source in the center of the array. The output from this step was feed into L1 regularized Least Square (LS) method for inferring the room geometry.

Dokmanic et al in [5] apply the properties of Euclidean Distance Matrices (EDMs) and Multidimensional Scaling (MDS) to iteratively find the room geometry in general 3D case with a single source and five randomly placed microphones. Adopted the same usage of EDM's property, Jager et al [6] further recast the echo labeling problem into finding the Maximum Independent Set (MIS) of a graph. This approach gave the same accuracy as Dokmanic's method at a lower computational complexity using at least 5 microphones and 2 sources. Nonetheless, both [5] and [6] will be unmanageable when the number of microphones increases. Finally, Coutino et al [7] succeed to reduce the computational complexity while maintained the sub-optimal performance in solving the echo labeling problem by implementing a subspace filtering method followed by a greedy-based rank constraint of an EDM.

# 3 Hybrid Method for Room Geometry Estimation Based on The Graph Based and The Subspace-Greedy Approach

This chapter introduces the improved solution of the acoustic echo sorting problem for a shoe box shaped room. The improved solution is based on the methods proposed by Jager [6] and Coutino [7]. The method of Jager and Coutino are called the graph based method and the subspace-greedy method respectively. Fig. 3.1 depicts the sequence of each method for estimating the room geometry.



$\tilde{D} \in \mathbb{R}^{M \times N^M}$
Echo labeling Solver
Pre –Filtering Stage
Rank Filter Threshold : $\varepsilon$
Echo Combination Candidate $\tilde{D}_{C_\varepsilon}$
Refinement Stage
Graph based Finding MIS
Maximum independent sets $S^G_{max}$
Labeled Distance Data $\Delta = [S \;\; E_1 \;\; E_2]$
Mics and Sources Localization
Pollefeys
Estimated Mic and source position $\hat{r}$ , $\hat{s}$
Procrustes for finding the best estimation of $\hat{s}$
Image source Position
$\tilde{D} \in \mathbb{R}^{M \times N^M}$
Echo labeling Solver
Pre –Filtering Stage
Subspace Filter Threshold : $\kappa$
Echo Combination Candidate $\tilde{D}_C$
Refinement Stage
Greedy based Rank criterion Threshold : $\varepsilon$
Estimated Distance Data $\hat{D}$
Sources Localization
Least squares $\hat{S} = (R^T)^+\hat{D}$
Estimated source position $\hat{s}$
Image source Position

(a) Block diagram of the graph based method (b) Block diagram of the subspace greedy method

Figure 3.1: Block diagram of the graph and the subspace-greedy methods

From Fig 3.1 we can see that we can further divide the echo labeling solver stage into two steps. The first step is the pre-filtering step and the second step is the refinement step. From the graph based method and the subspace greedy methods we are provided with some options of the technique for the echo labeling solver and the source localization procedure as listed below:

1. **Echo labeling solver**

   (a) Pre-filtering stage

   - Rank filtering

        - Subspace filtering

    (b) Refinement stage
        - Graph (finding a maximum independent set)
        - Greedy based with rank criterion.

2. **Image source localization method**
    - Pollefeys method
    - Least squares

Based on those options, we would examine some possible combinations that can be constructed to find a suitable combination with higher accuracy and faster computation time for estimating the room geometry. The detail of the combinations will be defined in Section 3.6.

## 3.1 Detailed Overview of Acoustic Echo Labeling Problem

As explained in the previous chapter, section 2.1, we face the echo ambiguity problem when determining the room boundaries (four walls, a floor, and a ceiling). Before we deal with the methods for solving the ambiguity problem, we initiate some assumptions that must hold :

1. **A correct estimation of the RIR**
   From RIR data, it is assumed that the correct peak related to the TOA is always available and extracted properly.

2. **Synchronization**
   The microphones and sources are synchronized so that the TOAs are absolute and directly correspond to distances.

3. **Known Microphones Position**
   The relative microphones position are known a priori.

Converting the TOAs into the squared distances and keeping those assumptions, then echo labeling problem can be defined as :

- ***Echo Labeling Problem***
  From an unlabeled squared distance matrix $\mathcal{D}$ between the $M$ microphones and the $N$ sources, it is required to find a correct labeled squared distance matrix $\mathbf{D} \in \mathbb{R}^{M \times N}$ which the $n$-th column comprises the squared distances between the $M$ microphones and the $n$-th image source $\forall\ n = 1, ..., N$. Note that each column of $\mathbf{D}$ is unique so the columns in $\mathbf{D}$ must not share elements in common.

To demonstrate how to handle this problem, let us assume that after extracting the RIR we have the following unlabeled squared distance matrix

$$\mathcal{D} = \{d_{mn}\} \forall\ m = 1, \cdots, M \text{ and } n = 1, \cdots, N\ . \tag{3.1}$$

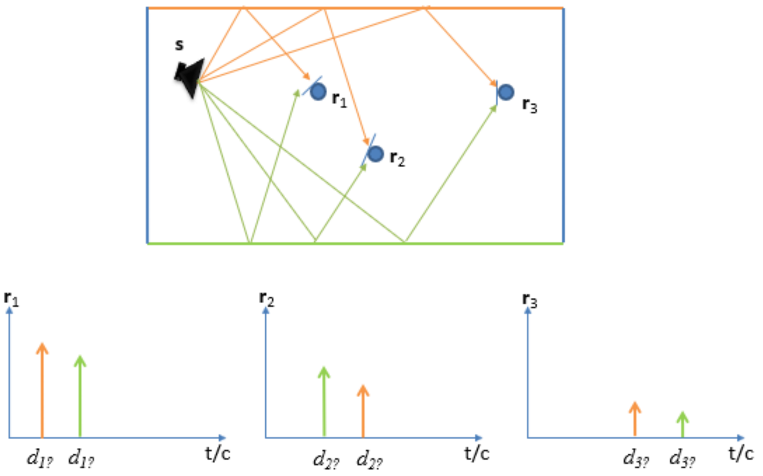


Figure 3.2: Ambiguity in the echoes that received by the microphones

Each $d_{mn}$ resembles the squared distance between the $m$-th microphone and the $n$-th image source. As the order of $n$ is unsorted ,i.e., the image source responsible for the echo is obscure, all possible echoes combination have to be generated and examined in order to find the correct group of echoes at different microphones that belongs to the same image source. This process produces a big $M \times N^M$ unlabeled squared distance matrix, $\tilde{\mathbf{D}}$.

The number of columns in $\tilde{\mathbf{D}}$ is $N^M$. Figure 3.2 shows a small example when the number of microphones is 3 and the number of image sources (walls) is 2. Factually, the three microphones have different echo arrival sequence so we do not know the precise order of echoes received in the microphones correspond to the green or the orange wall. Hence, all 8 ($2^3$)possible combinations of squared distance ($\tilde{\mathbf{D}}$) must be inspected to obtain the correct echo combination,

$$\tilde{\mathbf{D}} = \begin{bmatrix} d_{11} & d_{11} & d_{11} & d_{11} & d_{12} & d_{12} & d_{12} & d_{12} \\ d_{21} & d_{21} & d_{22} & d_{22} & d_{21} & d_{21} & d_{22} & d_{22} \\ d_{31} & d_{32} & d_{31} & d_{32} & d_{31} & d_{32} & d_{31} & d_{32} \end{bmatrix} \in \mathbb{R}^{M \times N^M}. \tag{3.2}$$

This condition reveals that the number of columns in $\tilde{\mathbf{D}}$ will increase exponentially as the number of microphones grows. Thus an algorithm that can reduce our column search space is necessary. We call this initial step a pre-filtering step. In the presence of noise, an additional step to refine the previous step's output is unavoidable since the pre-filtering step cannot directly give us the correct $\mathbf{D} \in \mathbb{R}^{M \times N}$ matrix.

## 3.2 Preliminary Equations and Relations

In this section the fundamental equation and relation will be featured. Assume we have a shoe box shaped room with $M$ microphones and $N$ sources randomly distributed inside a room with position in Cartesian coordinate defined by $\mathbf{r}_m = [x_m, y_m, z_m]^T \in \mathbb{R}^3$

and $\mathbf{s}_n = [X_n, Y_n, Z_n]^T \in \mathbb{R}^3$ respectively. The squared distance between the $(m,n)$-th microphone and source pair can be formulated as :

$$d_{m,n} = (x_m - X_n)^2 + (y_m - Y_n)^2 + (z_m - Z_n)^2 \, . \tag{3.3}$$

Expressed in vector notation, Eq. A.1 becomes

$$d_{m,n} = \mathbf{R}_m^T \mathbf{S}_n \, , \tag{3.4}$$

where

$$\begin{aligned} \mathbf{R}_m &= [\mathbf{r}_m^T \mathbf{r}_m \;\; -2x_m \;\; -2y_m \;\; -2z_m \;\; 1]^T \in \mathbb{R}^{5\times 1} \, , & (3.5) \\ \mathbf{S}_n &= [1 \; X_n \; Y_n \; Z_n \; \mathbf{s}_n^T \mathbf{s}_n] \in \mathbb{R}^{5\times 1} \, . & (3.6) \end{aligned}$$

Stacking up all the $(m,n)$ pair squared distance $d_{m,n}$ yields a squared distance matrix $\mathbf{D} \in \mathbb{R}^{M\times N}$ and Eq. A.2 can be expressed in a matrix form as

$$\mathbf{R}^T \mathbf{S} = \mathbf{D} \in \mathbb{R}^{M\times N} \, , \tag{3.7}$$

where $\mathbf{R} = [\mathbf{R}_1, ..., \mathbf{R}_M]$ and $\mathbf{S} = [\mathbf{S}_1, ..., \mathbf{S}_N]$ are the microphone and source position matrices according to Eq. 3.5 and 3.6.

## 3.3 Pre-filtering Step

The input for this step is all possible column combinations of the squared distance data ($\tilde{\mathbf{D}} \in \mathbb{R}^{M\times N^M}$). The first state-of-the-art method proposed by Jager et al. in [6] used a method based on the EDM rank property of the squared distance matrix while Coutino et al. in [7] exploited the subspace relation of the squared distance data and a microphone position matrix ($\mathbf{R}$).

### 3.3.1 Rank Filtering

Assuming the knowledge of the microphones position $(\mathbf{r}_1, \cdots, \mathbf{r}_m)$, an EDM ($\mathbf{E}$) can be built using the squared Euclidean distance between all pairs of the microphones $(d_{r_i r_j})$,

$$\mathbf{E} = \begin{bmatrix} d_{r_1 r_1} & d_{r_1 r_2} & \cdots & d_{r_1 r_M} \\ d_{r_2 r_1} & d_{r_2 r_2} & \cdots & d_{r_2 r_M} \\ \vdots & \vdots & \ddots & \vdots \\ d_{r_M r_1} & d_{r_M r_2} & \cdots & d_{r_M r_M} \end{bmatrix} \in \mathbb{R}^{M\times M} \, . \tag{3.8}$$

According to Theorem 1 in Section 2.1.4, the rank of an EDM, $\mathbf{E}$ with affdim$(\mathbf{E}) = 3$ is at most 5. By appending each column of $\tilde{\mathbf{D}}$ to $\mathbf{E}$ one column at a time, we augment the $\mathbf{E}$ and form an $\tilde{\mathbf{E}}$,

$$\tilde{\mathbf{E}} = \begin{bmatrix} \mathbf{E} & \tilde{\mathbf{D}}_c \\ \tilde{\mathbf{D}}_c^T & 0 \end{bmatrix} \in \mathbb{R}^{(M+1)\times(M+1)} \, , \tag{3.9}$$

$\tilde{\mathbf{D}}_c \in \mathbb{R}^{M\times 1}$ denotes the $c$-th column of $\tilde{\mathbf{D}}$. Then we impose the rank constraint on $\tilde{\mathbf{E}}$ to check whether each column of $\tilde{\mathbf{D}}$ is feasible as a true echo combination. If $\mathbf{E}$ is augmented with the correct echoes combination, rank($\tilde{\mathbf{E}}$) should be $\leq 5$. However, in a real situation where erroneous TOA estimation is inevitable, this rank test always fails and we get an empty set. To overcome this problem, Jager et al. in [6] proposed a tolerance called $\varepsilon$. Instead of rank($\tilde{\mathbf{E}}$), now we will consider:

$$\text{rank}(\tilde{\mathbf{E}}, \varepsilon) = \min_{||\tilde{\mathbf{E}}-\mathbf{X}||_2 \leq \varepsilon} \text{rank}(\mathbf{X}) \,. \tag{3.10}$$

This condition filters out singular values from the SVD of $\tilde{\mathbf{E}}$ which is greater than $\varepsilon$ and generates a set of column indexes given by

$$\mathcal{C}_\varepsilon = \{c : \text{rank}(\tilde{\mathbf{E}}, \varepsilon) \leq 5\} \,. \tag{3.11}$$

The output of this step is a matrix of echo combination candidates which contains the column of $\tilde{\mathbf{D}}$ listed in $\mathcal{C}_\varepsilon$ (Eq.3.12) :

$$\tilde{\mathbf{D}}_{\mathcal{C}_\varepsilon} \in \mathbb{R}^{M \times |\mathcal{C}_\varepsilon|} \,. \tag{3.12}$$

In the case of noiseless measurement data, $\tilde{\mathbf{D}}_{\mathcal{C}_\varepsilon}$ provides exactly $N$ columns correspond to the true echo combination. Unfortunately, in practice this is unrealistic. Thus $|\mathcal{C}_\varepsilon| \gg N$ in practice.

### 3.3.2 Subspace Based Method

The subspace based method is an alternative method for solving echo labeling problem presented by [7] which has a similar function as rank filtering. It is able to perfectly select the correct columns of $\tilde{\mathbf{D}}$ under a noise free condition.

Recall the microphone position matrix ($\mathbf{R}$) from Eq. A.5, the singular value decomposition (SVD) of this matrix is :

$$\mathbf{R} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T \,. \tag{3.13}$$

From here we can calculate a projection matrix which projects any vector in the null space of $\mathbf{R}$ ($ker(\mathbf{R})$) denoted as $\Pi_{\mathcal{N}(\mathbf{R})}$,

$$\Pi_{\mathcal{N}(\mathbf{R})} = \mathbf{I}_M - \tilde{\mathbf{V}}\tilde{\mathbf{V}}^{\mathbf{T}} \,, \tag{3.14}$$

where $\tilde{\mathbf{V}} \in \mathbb{R}^{M \times 5}$ is the economy size $\mathbf{V}$ matrix from the SVD of $\mathbf{R}$. Applying $\Pi_{\mathcal{N}(\mathbf{R})}$ to $\mathbf{D}$ and $\mathbf{R}$ it can be shown that

$$\Pi_{\mathcal{N}(\mathbf{R})}\mathbf{D} = \Pi_{\mathcal{N}(\mathbf{R})}\mathbf{R}^T\mathbf{S} = 0 \,, \tag{3.15}$$

which can be employed to estimate $\mathbf{D}$ from $\tilde{\mathbf{D}}$. This projection matrix has an interesting property

$$\|\Pi_{\mathcal{N}(\mathbf{R})}\|_2 = 1 \,, \tag{3.16}$$

which implies that there is no amplification errors, i.e.,

$$\begin{aligned} \|\Pi_{\mathcal{N}(\mathbf{R})}\mathbf{D} + \mathbf{N}\|_2 &= \|\Pi_{\mathcal{N}(\mathbf{R})}(\mathbf{R}^T\mathbf{S} + \mathbf{N})\|_2 && (3.17) \\ &= \|\Pi_{\mathcal{N}(\mathbf{R})}\mathbf{N}\|_2 && (3.18) \\ &\leq \|\mathbf{N}\|_2 && (3.19) \\ &= \sigma_{max}(\mathbf{N}) \,, && (3.20) \end{aligned}$$

where $\sigma_{max}(\mathbf{N})$ is the maximum singular value of the noise matrix $\mathbf{N}$. This fact makes the application of projection matrix useful when the elements of $\tilde{\mathbf{D}}$ are contaminated with noise.

To appreciate how the subspace filtering works, let us define a function for a projection of $\tilde{\mathbf{D}}$ into the nullspace of $\mathbf{R}$:

$$f(c) = \|\Pi_{\mathcal{N}(\mathbf{R})}\tilde{\mathbf{d}}_c\|_2^2 \quad \forall\, c\, \in [1, \cdots, N^M]\,, \tag{3.21}$$

where $\tilde{\mathbf{d}}_c$ is the $c$-th column vector of $\tilde{\mathbf{D}}$. Using the property in Eq. 3.15, we can select the subset of feasible columns

$$\mathcal{C} = \{c : f(c) = 0\}\,, \tag{3.22}$$

and the candidate echoes combinations which retain the column specified by $\mathcal{C}$ is given by

$$\tilde{\mathbf{D}}_{\mathcal{C}} \in \mathbb{R}^{M\times|\mathcal{C}|}\,. \tag{3.23}$$

However, in the noisy condition the set $\mathcal{C}$ in Eq.3.22 turns empty. To deal with this issue, Coutino et al in [7] introduced an upper bound ($\kappa_c$), then Eq.3.22 is rewritten as

$$\mathcal{C} = \{c : f(c) \leq \kappa_c\}\,, \tag{3.24}$$

$$\kappa_c \;= 4\,\gamma\,\sigma_N^2\,\|\tilde{\mathbf{d}}_{\mathbf{c}}^{\circ\frac{1}{2}}\|_2^2,\;\gamma \geq 1\,. \tag{3.25}$$

Complete derivation of $\kappa_c$ can be found in [7]. The accuracy of the method for estimating the TOA is considered to be known as well as the noise power $\sigma_N^2$. For simplicity, the authors of [7] presume that all columns in $\tilde{\mathbf{D}}$ are subject to the same noise level $\sigma_N^2$. The upper bound $\kappa_c$ raises the same problem about overestimation of $|\mathcal{C}|$. Therefore, the following section will explain the next step to refine this intermediate result.

## 3.4 Refinement Step

Both pre-filtering steps always end up with an overestimated number of squared distance columns ($|\mathcal{C}_\varepsilon|$ and $|\mathcal{C}| \gg N$). Consequently, another step is required to select the appropriate columns. There are two possible methods that can be implemented: $(i)$ the graph based method, $(ii)$ the greedy based rank criterion, as we will see in more detail in Section 3.4.1 and 3.4.2. The basic idea behind the graph and the greedy method is the same, i.e., the columns in $\mathbf{D}$ must not share elements in common.

### 3.4.1 Graph Based (Maximum Independent Set)

The input of this step is $\tilde{\mathbf{D}}_{\mathcal{C}_\varepsilon}$ with $|\mathcal{C}_\varepsilon| \gg N$. To find the correct echo combination, Jager et al.[6] formulated the problem as a graph problem of finding the maximum independent sets (MISs) from $\tilde{\mathbf{D}}_{\mathcal{C}_\varepsilon}$. This method relies on the fact that the Euclidean squared distances between the image sources and microphones are unique.

First, each column in $\tilde{\mathbf{D}}_{\mathcal{C}_\varepsilon}$ is modeled as a node ($V$) in an undirected graph $G(V, E)$. An edge ($E$) between two nodes will be defined if their corresponding columns share

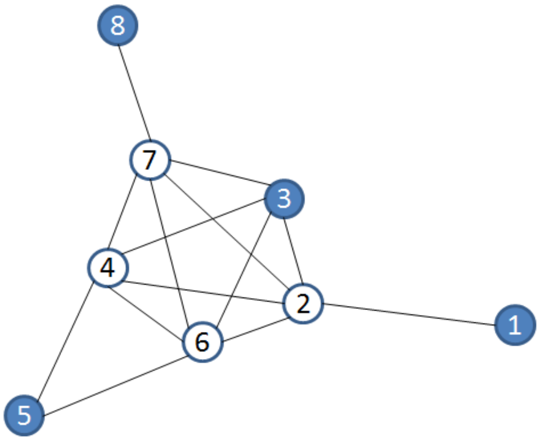


Figure 3.3: Echo combinations in $\tilde{\mathbf{D}}_{\mathcal{C}_\varepsilon}$ as nodes in a graph. The set of blue nodes is the (maximum) independent set.

one or more elements in common. Then, the algorithm will search a subset of all echo combinations (nodes) that do not share connections in the graph. Suppose in 2D there are three microphones and four walls, and assume that $\tilde{\mathbf{D}}_{\mathcal{C}_\varepsilon}$ is given by :

$$\tilde{\mathbf{D}}_{\mathcal{C}_\varepsilon} = \begin{bmatrix} 8 & 9 & 9 & 9 & 10 & 10 & 12 & 12 \\ 18 & 18 & 19 & 20 & 20 & 19 & 22 & 22 \\ 28 & 29 & 29 & 29 & 31 & 29 & 29 & 32 \end{bmatrix}. \tag{3.26}$$

From Eq. 3.26, we need to find a set with four columns ($c_i \in \tilde{\mathbf{D}}_{\mathcal{C}_\varepsilon}$) that represent the correct echo combination. By inspection, it is clearly seen that an independent set of $(c1, c3, c5, c8)$ is the required MIS since this is the only set with four columns that do not have common elements. Figure 3.3 shows the graph representation of Eq. 3.26 where the blue nodes describe the independent set.

In graph theory, a set of nodes with no adjacency between all possible node pair is called an independent set. There can be many independent sets in a graph. By definition, each subset of an independent set is also an independent set. A maximal independent set is a set such that adding any other node forces the set to have an edge and makes the set unqualified as an independent set. The size of an independent set is the number of nodes in that set. An independent set with the largest possible size in a graph is called a maximum independent set (MIS). Thus, the graph based algorithm tries to find the MIS ($\mathcal{S}^G_{max}$) from $G(V, E)$. In general, there can be more than one MIS ($|\mathcal{S}^G_{max}| > 1$) in a graph and we need to find all of them. One way to do this is by listing all maximal independent sets which leads to an NP-hard problem. As a result, an additional step is required to choose the correct set which corresponds to the correct echoes combination.

### 3.4.2 Greedy with Rank Criterion

Coutino et al. in [7] came with the greedy strategy which adopts the rank test procedure of the augmented EDM, $\tilde{\mathbf{E}}$ (Section 3.3.1) and the key observation that two columns from $\mathbf{D}$ must not share elements in common came with the greedy strategy. This algorithm acts as a substitute for the graph based method to avoid the NP-hard problem of an exhaustive search through all maximal independent sets.

Following the scheme in [7], the subspace pre-filtering step produces the truncated $\tilde{\mathbf{D}}$ matrix, i.e. $\tilde{\mathbf{D}}_{\mathcal{C}}$ which already sorted in an ascending fashion based on the functional value ($f(c)$) defined in Eq.3.21. This matrix acts as an input for the greedy procedure. Subsequently, the greedy procedure will do the rank test (Section 3.3.1) using the first column of $\tilde{\mathbf{D}}_{\mathcal{C}}$, if this column passes the test, then it will be kept and the algorithm will continue to do the rank test using the second column. To avoid common elements, an additional test which compares the elements of the second column with the column that already kept is also applied. If the second column passes both tests, it will be saved. If not, then the algorithm picks the next in line column of $\tilde{\mathbf{D}}_{\mathcal{C}}$ and redo the same procedure. This process ends when the number of columns that are saved equals to $N$ (the number of walls). In this way, the greedy procedure always delivers a unique set of the squared Euclidean distance $\hat{\mathbf{D}} \in \mathbb{R}^{M \times N}$.

## 3.5 Source Position Localization

The output of the graph based method is the maximum independent sets ($\mathcal{S}^G_{max}$). Each maximum independent set contains echo combinations that do not have elements in common. If $\mathcal{S}^G_{max}$ consists of one set only, this set is the correct echo combination of which we can infer the room geometry using simple least squares (Eq. 3.27). However, if $\mathcal{S}^G_{max}$ contains more than one independent set, an additional method to decide the correct set is required. On the other hand, the output of the greedy procedure always gives a unique set so we can apply the least squares to estimate the (image) sources position, i.e.,

$$\hat{\mathbf{S}} = (\mathbf{R}^T)^{\dagger}\mathbf{D}\,, \tag{3.27}$$

$\dagger$ means the pseudoinverse. To choose the correct set in case $|\mathcal{S}^G_{max}| > 1$, Jager et al. in [6] utilized the source localization algorithm proposed by Pollefeys [15]. Once the image source positions are obtained, the room geometry can be reconstructed using straightforward geometrical methods as explained in Chapter 2 (Section 2.1.3).

### 3.5.1 Pollefey's method

The Pollefeys' method is able to localize both microphones and sources position up to unitary transforms (rotation, reflection) and the translation given the correct labeled set of distance data. This method works based on the rank-5 factorization of $\mathbf{D}$. The detailed explanation of this method can be found in Appendix A. Holding the prior knowledge of microphone location, we can use it as a tool to check whether the source localization results are correct or not by comparing the estimated microphones' location with the true microphone location using Procrustes analysis [16].

Pollefeys' method required at least ten sources and five microphones. Using the image source method which models the reflections as virtual sources, we need to have at least two real sources that yield 12 image sources in the case of a 3D shoe box room to fulfill the requirement. Thus, Pollefeys' method forces us to provide at least two non-collocated real sources ($N \geq 2$). Given $N = 2$ , the input data for the Pollefeys' method can be constructed as

$$\Delta = [\mathbf{D}_S \; \mathbf{E}_1 \; \mathbf{E}_2] \;, \tag{3.28}$$

where $\mathbf{D}_S \in \mathbb{R}^{M \times N}$ is the squared distance matrix between $M$ microphones to the $N$ real sources, while $\mathbf{E}_1 \in \mathcal{S}_{max}^{G_1}$ and $\mathbf{E}_2 \in \mathcal{S}_{max}^{G_2}$ are a subset of the MIS from the first real source and the second real source respectively. Both $\mathbf{E}_1$ and $\mathbf{E}_2$ are $\in \mathbb{R}^{M \times 6}$ which represent the squared distance matrix between the $M$ microphones and six virtual sources with respect to the first and the second real sources. In case of $N > 2$, all combinations from the sets in $\{\mathcal{S}_{max}^{G_1}, \mathcal{S}_{max}^{G_2}, \cdots, \mathcal{S}_{max}^{G_N}\}$ have to be created and acted as the Pollefeys input

$$\Delta = [\mathbf{D}_S \; \mathbf{E}_i \; \mathbf{E}_j] \;, \tag{3.29}$$

where $1 \leq i, \; j \leq N \;, i \neq j$, so in total we can make $\binom{N}{2}$ combinations.

## 3.6 Comparison of The Graph Based Method and The Subspace-Greedy Method

Recall Figure 3.1, Table 3.1 provides the comparison of the computational complexity for the graph based and the subspace greedy method. It reveals that the graph based method has a higher computational cost than the subspace-greedy approach. This condition is mainly caused by the rank filtering step since we need to redo the SVD $N^M$ times for each augmented EDM. The next contributor to the slowness of the graph based method is the graph technique for finding the MIS. If the threshold ($\epsilon$) in the rank filtering step is quite loose, a lot of columns from $\tilde{\mathbf{D}}$ are passed, and the graph algorithm takes a long time for discovering the MIS. The last contributor comes from the Pollefeys' algorithm particularly when the cardinality of the MIS is large (more pairs are needed to be checked).

| **Graph based** | | **Subspace - greedy** | |
|---|---|---|---|
| **Step** | **Complexity** | **Step** | **Complexity** |
| Rank filter | $N^M \mathcal{O}((M+1)^3)$ | Subspace filter | $N^M \mathcal{O}(M^2)$ |
| graph (MIS) | $\mathcal{O}(2^{0.276\lvert\mathcal{C}_\varepsilon\rvert})$ | greedy(rank) | $\lvert\mathcal{C}\rvert \mathcal{O}((M+1)^3)$ |
| Pollefeys + Procrustes | $\left[\prod_{i=1}^{N} \lvert\mathcal{S}_{max}^{G_i}\right] \mathcal{O}(49MN^2)$ | least squares | $\mathcal{O}(NM^2)$ |

Table 3.1: Computational complexity comparison of each step in the graph based and the subspace greedy methods [1].

Table 3.2 summarizes the advantage(s) and disadvantage(s) of both the graph based and the subspace greedy methods. In this chapter, we aim to improve the drawback of the graph based method. First, we eliminate the minimum source requirement of the graph based method by exchanging the Pollefeys' algorithm with the least squares, after that we compare the outcome of the Pollefeys and the least squares. Furthermore, we want to avoid solving the graph problem of finding the MIS which is an NP-hard

| **Method** | **Plus** | **Minus** |
|---|---|---|
| **Graph based** | -have the optimal accuracy | -high computational cost<br>-must provide at least two real sources |
| **Subspace-greedy** | -low computational cost<br>-gives a unique $\hat{\mathbf{D}}$<br>-works with a single source | -give a suboptimal performance |

Table 3.2: The plus and minus of graph based and subspace-greedy method.

problem in the refinement step of the echo labeling solver so we follow the greedy procedure that was described in Section 3.4.2 but we remove the rank test part because the graph based method has already applied the rank test in the pre-filtering stage. The next section will provide the simulation results of these actions which demonstrate the first contribution of this thesis.

## 3.7 Experimental Results

The results from a set of simulations are displayed to evaluate the proposed changes in the graph based method. The performance of the result is assessed in terms of accuracy and computational speed against the number of microphones, the number of sources, and the noise standard deviation of the distance data. For each evaluation, a set of 100 simulations were performed using a synthetic distance data between the microphones and the image sources that are perturbed with the white Gaussian noise $\sim \mathcal{N}(0, \sigma^2)$ to simulate uncertainties in the TOA estimation. The sources and microphones are randomly placed inside a shoe box shaped room with a constant dimension ($8m \times 6m \times 5m$). The simulations were run in Matlab.

To quantify the accuracy of each method, the Frobenius norm of the difference between the estimated and the true vertices position in 2D was used (Eq. 3.30),

$$\text{Error}(\hat{\boldsymbol{\theta}}) = \frac{1}{N} \sum_i \|\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i\|_F \ , \ i = 1, 2, \cdots, N \ , \tag{3.30}$$

where $\boldsymbol{\theta}_i$ and $\hat{\boldsymbol{\theta}}_i$ represents the true room vertices position and the estimated vertices position for the $i$-th experiment respectively.

### 3.7.1 Comparison between Pollefeys and the least squares as the image source localization method of the graph based method

Figure 3.4 illustrated the block diagram of the experiment that will be done to compare Pollefeys and the least square method which serve as the source localization procedure of the graph based method. The outcome of the echo labeling solver block is an MIS ($\mathcal{S}^G_{max}$) but as already stated earlier in Section 3.4.1, the graph technique does not guarantee the uniqueness of the MIS ($|\mathcal{S}^G_{max}| > 1$). In the original version of the graph based method [6], this non unique MIS problem is handled by always taking the

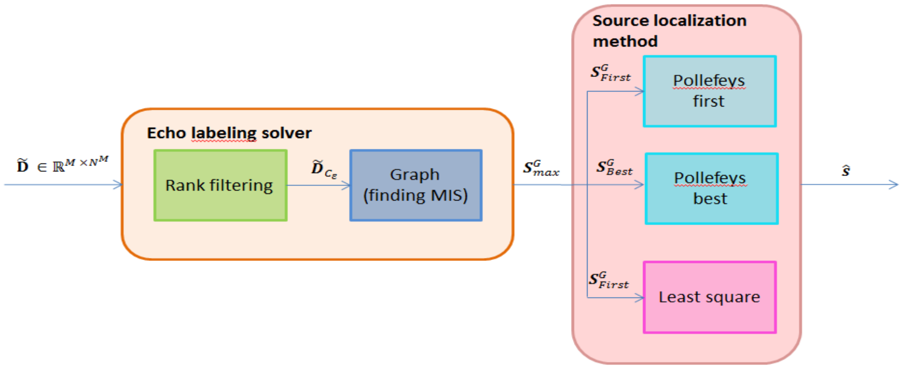


Figure 3.4: The block diagram for the first experiment

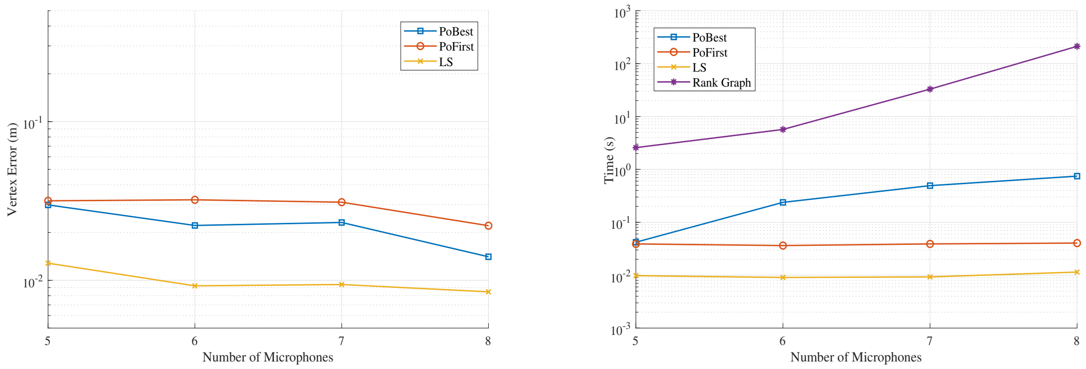
best paired MIS ($\mathbf{E}_i$ , $\mathbf{E}_j$), i.e., the paired set that produces the smallest microphone reconstruction error. However, for the least square method if $|\mathcal{S}^G_{max}| > 1$ we cannot apply the same checking procedure since the least squares directly gives the source's position. Hence in the experiment, the least square method will always take the first set of $\mathcal{S}^G_{max}$ as its input. To make a fair comparison We also check the performance of the Pollefeys method if the first MIS is employed as its input. Following Figure 3.4, we will compare the performance of these three methods: Pollefey first (PoFirst), Pollefey best (PoBest), and least squares in terms of accuracy and computational time. The experiments were done with the following setup :

1. The number of microphones ($N$) is varied, $N = 6$, $\sigma = 0.001m$.
2. The number of sources ($M$)is varied, $M = 6$, $\sigma = 0.001m$.
3. The noise standard deviation ($\sigma$) is varied, $M = 6$, $N = 6$.

#### 3.7.1.1 Variation in the number of microphones

For these simulations, we consider the case of $N = 6$ sources distributed randomly inside a room and the noise standard deviation ($\sigma$) = 1 mm. The number of microphones is varied from five to eight. The minimum number of microphones is set to five to follow the Pollefeys' method requirement. In [6], the author stated that the computational time of the graph based method become intractable when $M > 7$ due to the echo labeling solver part. Since here we deal with the source localization procedure, we are curious how it will perform in case $M = 8$.

The graph in Fig. 3.5a suggests that the vertex estimation error decreases as the number of microphones increases. This behavior is logical since adding more microphones assures that both the matrix $\mathbf{R}$ and the EDM of microphone paired distance ($\mathbf{E}$) have rank 5. Consequently, the accuracy of the three methods is improved. Moreover, Fig. 3.5a clearly shows that the accuracy of PoFirst (red line) is slightly worse than PoBest with the difference ±0.01$m$ while the least square gives the lowest vertex estimation error.



(a) Vertex estimation error VS Number of microphones

(b) Computational Time VS number of microphones

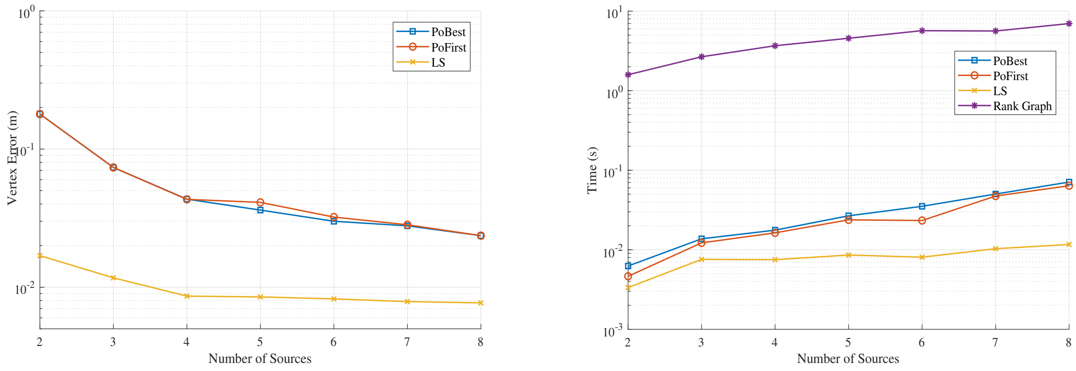
Figure 3.5: The number of microphones varied, $N = 6$, $\sigma = 0.001$

In terms of computational cost, raising the number of microphones has no significant effect for PoFirst and the least square (red and yellow line in Fig. 3.5b), whereas for PoBest the increment of $M$ increased the computational time by 0.86 second (from $M = 5$ to $M = 8$). Fig.3.5b also supports the superiority of the least square method against Pollefeys. Unfortunately, the reduced computational time that we gain by substituting Pollefeys with the least squares in the source localization procedure did not play an important role in minimizing the total computational time of the graph based method because most of the computational time is occupied by the echo labeling solver part (rank-graph) as illustrated by the purple line in Fig. 3.5b. The reason behind this condition is the growth of the number of columns in $\tilde{\mathbf{D}}$ that is needed to be checked increases exponentially.

#### 3.7.1.2 Variation in the number of Sources

The number of microphones ($M$) was set to 6 and the noise standard deviation ($\sigma$) = 1 mm. The results are depicted in Figure 3.6.

Figure 3.6a shows that both Pollefey and least square vertex estimation error decrease as the number of sources increases. The noticeable decrease occurs when $N$ goes from 2 to 4. For Pollefeys, the gradual decrement still happens as $N$ grows. This condition occurs because by adding more sources, the number of paired sources will increase and will compensate the paired source which cannot work together. On the other hand, after $N = 4$ the least squares has given a stable behavior since the least square method does not depend on the paired combination, but on the pseudo inverse of the microphone position matrix ($\mathbf{R}$). The notable decrement for all methods when $N$ increases from 2 to 4 happens because the algorithm for estimating the vertex relies on minimum two wall points. If the noise is very bad, then the estimation of image source position corresponds to one of the sources will be wrong, then it will affect the vertex estimation result. However, when $N$ increases, the number of wall points will



(a) Vertex estimation error VS Number of sources (b) Computational Time VS number of sources

Figure 3.6: The number of sources varied

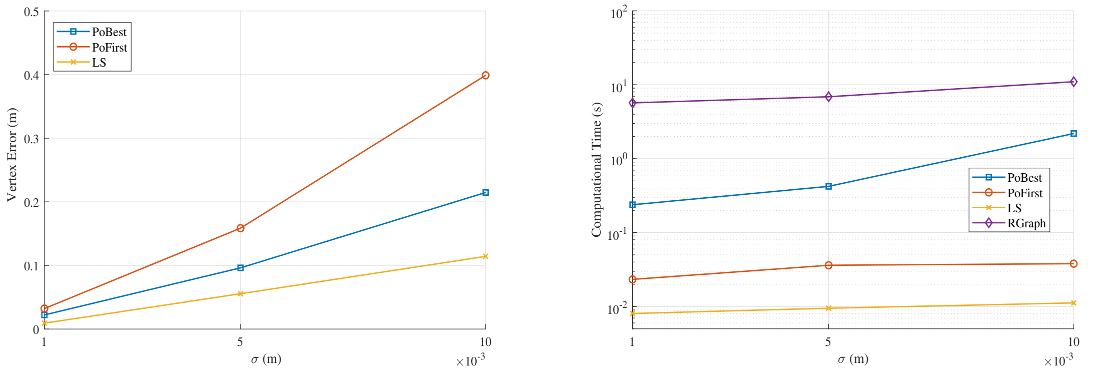
also increase. Thus, another wall points can remedy the bad point and help the vertex estimation result.

Figure 3.6 displays the superiority of the least square method compared to Pollefeys method in terms of accuracy and computational cost. Moreover, the accuracy that achieved by PoBest and PoFirst is the same. It means that in most of the cases, the first independent set corresponds to the true echo combination. The reason behind this phenomenon is that the columns in $\tilde{\mathbf{D}}$ have been sorted based on the sixth singular value of the augmented inter-microphones EDM in an ascending order before entering the graph method. This sorting makes the first MIS contains the squared distance columns that produce smallest 6-th singular of the augmented EDM.

Fig. 3.6b provides the same conclusion as in the previous section, the computational cost for least square always lower than the Pollefeys since the least square does not need to build a pair input combination also it just involves the inversion of a matrix and matrix multiplication. In this experiment, since most of the sources produce a unique MIS then the computational time between PoFirst and PoBest is almost overlap.

#### 3.7.1.3 Variation in The Distance Data Uncertainty

Intuitively, increasing the noise standard deviation will increase the vertex estimation error of the algorithm as depicted in Fig. 3.7. Note that in this case the accuracy of PoBest and PoFirst is diverged (the blue and red line in Fig. 3.7a) when the noise standard deviation upsurge. An interesting behavior is shown in the computational time curve (Figure 3.7b). The gap between the computational time of PoBest and the rank-graph filtering becomes closer as the noise increases because the number of non unique MIS from each source increases ($|\mathcal{S}^{G_m}_{max}| \gg 1, m = 1, 2, \cdots, M$), then a lot of paired source combinations have to be tested by PoBest.



(a) Vertex estimation error VS noise standard deviation (Graph)

(b) Computational Time VS noise standard deviation (Graph)

Figure 3.7: The noise standard deviation varied

### 3.7.2 Comparison of Graph vs Greedy as The Refinement Step of The Graph Based Method

In this part, we exchange the refinement technique of finding the MIS in the graph based method with the greedy approach. Furthermore, here we adopt the greedy procedure, but leave the rank criterion since the pre-filtering step of the graph based method already used the rank test to filter the column in $\tilde{\mathbf{D}}$. Thus, all the filtered columns ($\tilde{\mathbf{D}}_{\mathcal{C}_\varepsilon}$) have satisfied the rank criterion. Consequently, the greedy procedure just aims to find a set of six columns from $\tilde{\mathbf{D}}_{\mathcal{C}_\varepsilon}$ with no elements in common and produce a unique set of $\mathbf{D}$ so there is no difference between PoBest and PoFirst. The block diagram of the experiment that will be done is depicted in Fig. 3.8. It is obvious from Fig. 3.8 that we will have six lines in the final curve since we implement two echo labeling solver technique and three source localization method. The same set up as described in Section 3.7.1 was employed for the experiments.

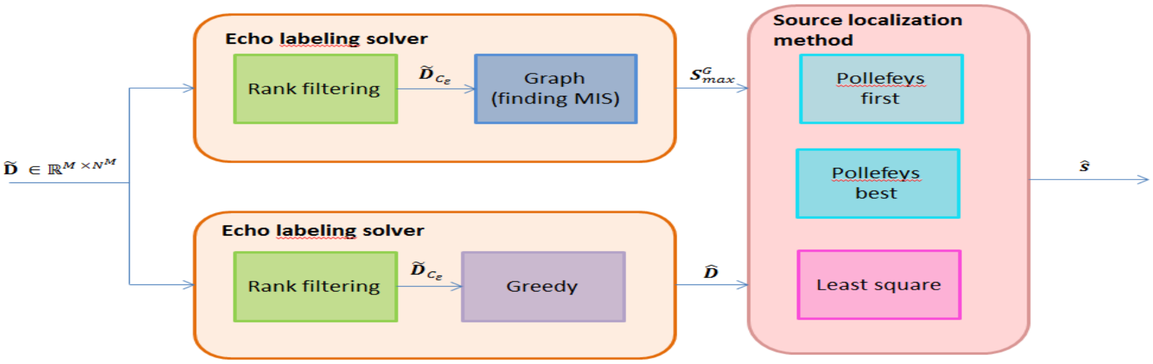
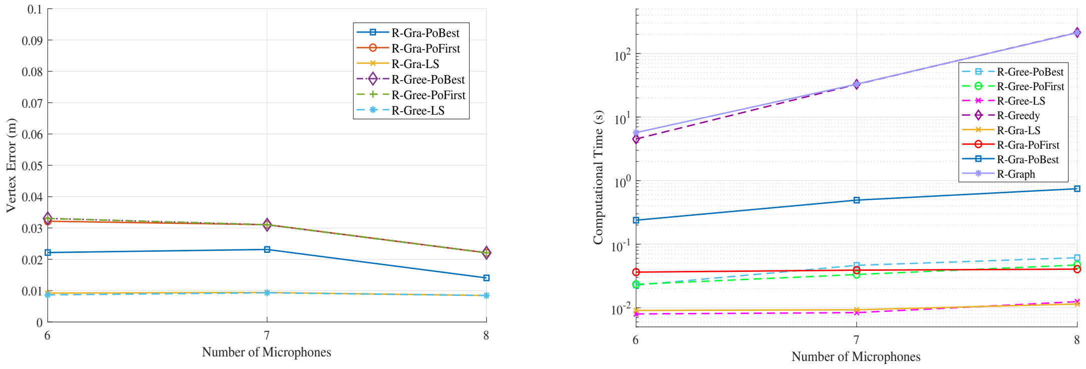


Figure 3.8: Block diagram of the second experiment.

In Fig. 3.9 the error and computation time of all methods are shown for the different number of microphones. At $M = 5$, the greedy procedure cannot perform well (most of the time it picks the wrong column) so we decided to start with M = 6.

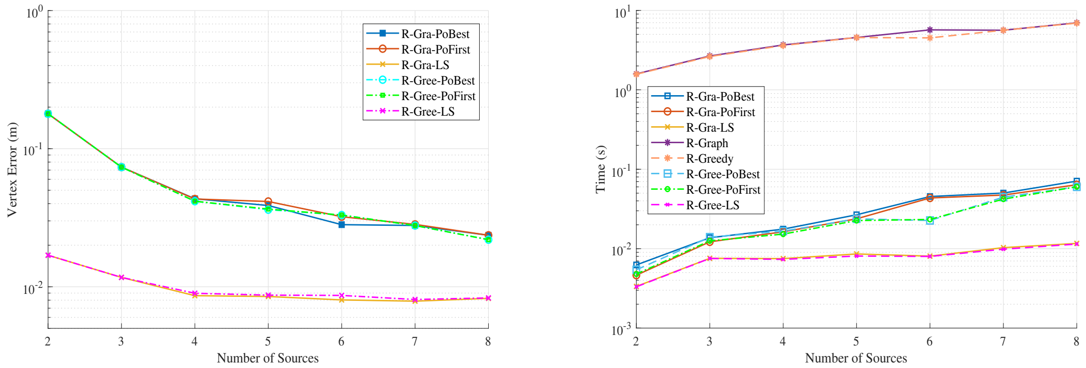


(a) Vertex estimation error VS number of microphones

(b) Computational Time VS number of microphones

Figure 3.9: Comparison of the greedy and the graph technique as the refinement step of the graph based method when $M$ is varied, $N = 6$ , $\sigma = 1mm$.

Starting with $M = 6$ the performance of both source localization methods after the greedy procedure (the dotted line) is almost the same as after the graph based procedure (the solid line).



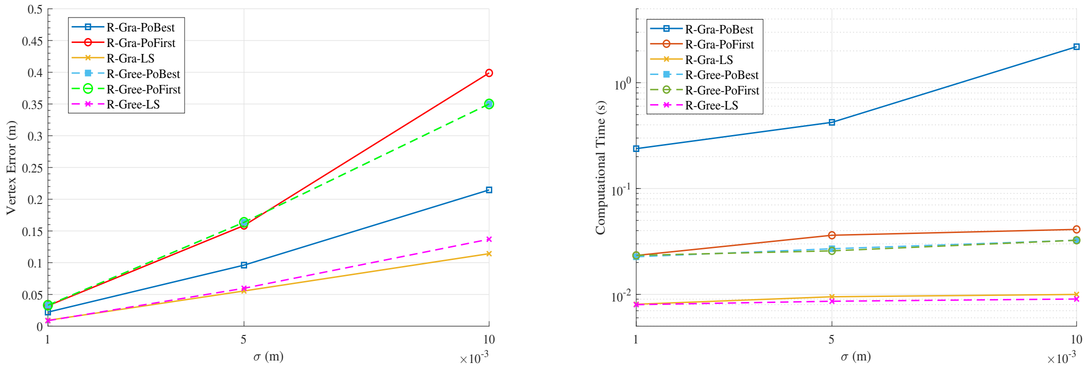
(a) Vertex estimation error VS number of sources

(b) Computational time VS number of sources

Figure 3.10: Comparison of the greedy and the graph technique as the refinement step of the graph based method when $N$ is varied, $M = 6$ , $\sigma = 1mm$.

For the computational time, the source localization techniques preceded by the greedy procedure are faster than the source localization techniques preceded by the graph method except for the least square method. The least square method gives the same computation time, regardless of which echo labeling solver was used. From Fig. 3.9b, we can see the time that is consumed for computing source localization method is smaller than the time needed for the echo labeling solver. This statement is also supported by Fig. 3.10b and 3.12.

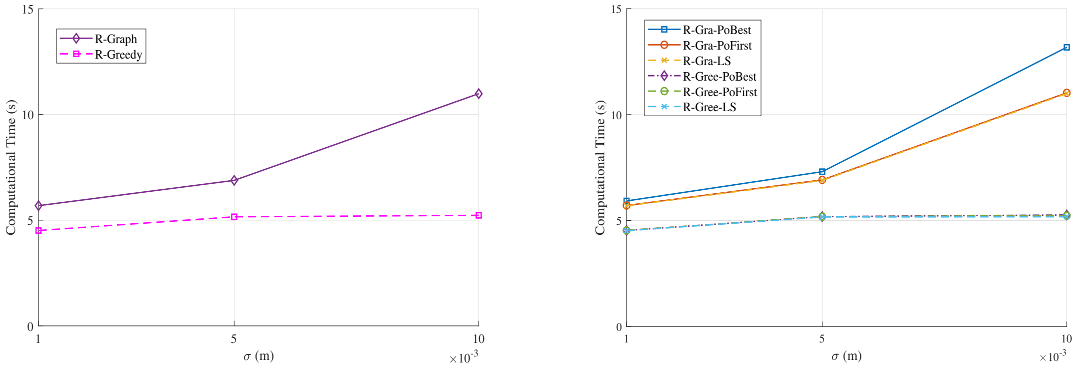
Fig. 3.12 indicates that most of the total computational time is occupied by the echo labeling solver part. As we can see that the curve in Fig. 3.12b has the same shape as the curve in Fig. 3.12a.



(a) Vertex estimation error VS noise standard deviation ($\sigma$)

(b) Computational Time VS noise standard deviation ($\sigma$)

Figure 3.11: Comparison of the greedy and graph method as the refinement step of the graph based method when $\sigma$ is varied, $M = 6\,, N = 6$.



(a) Computational time of echo labeling solver step (rank- graph and rank-greedy) when the noise standard deviation ($\sigma$) is varied.

(b) Computational time of the complete algorithm as described in Fig.3.8 against the noise standard deviation, $\sigma$

Figure 3.12: Computational time of echo labeling solver part (a) and the complete algorithm (b)in estimating the room geometry when $\sigma$ is varied.

## 3.8 Summary

From the previous section, some important points are derived :

1. Based on Jager's and Coutino's methods we can develop more combination methods of echo labeling solver and source localization procedure:

   (a) rank - graph - Pollefeys (the original version of graph based method)

   (b) rank - graph - least square

   (c) rank - greedy - Pollefeys

   (d) rank - greedy - least square

   (e) subspace - greedy - least square (the original version of the subspace greedy method)

   (f) subspace - greedy - Pollefeys

   (g) subspace - graph - least square

   (h) subspace - graph - Pollefeys

   In this chapter, our experiments are focused on the first four combinations since the last four combinations had been examined in [7].

2. From the first experiment which compares Pollefeys and the least squares (methods a and b), it can be concluded that the performance of the least square surpasses Pollefeys in both accuracy and computational time.

3. The second experiment reveals that the fourth method (d) in point 1 is the best combination which provides lower vertex estimation error although, in terms of computational time, the improvement is insignificant. Table 3.3 shows the result for the experiment in this chapter for $M = 6$, $N = 6$, $\sigma = 0.001$, and roomsize = [8 6 5].

| **Method** | **Vertex Estimation Error (m)** | **Computational Time (s)** |
|---|---|---|
| **Rank-Graph-PoBest** | 0.022 | 6.25 |
| **Rank-Graph-PoFirst** | 0.032 | 6.035 |
| **Rank-Graph-LS** | 0.0095 | 6.009 |
| **Rank-Greedy-Pollefeys** | 0.033 | 4.522 |
| **Rank-Greedy-LS** | 0.0094 | 4.509 |

Table 3.3: The final result of both experiment for $M = 6$, $N = 6$, $\sigma = 0.001$, and roomsize = [8 6 5]

# Implementation of the Hybrid Methods on The Sphere's Surface Microphones Constellation 4

The aim of this chapter is to check whether the earlier methods derived in Chapter 3 are feasible for estimating the room geometry with a new sphere's surface microphones constellation and to find a solution for problems that arise due to the close microphones arrangement.

## 4.1 Microphones Constellation

As mentioned before in Chapter 1, measuring the room geometry with movable and small microphones configurations is preferable especially for applications in robot navigation and virtual reality games. Driven by these reasons, we choose to put the microphones randomly on the surface of a small sphere with a radius that resembles the head of a person or a robot on average ($\pm 0.10m$). Figure 4.1 displays one possible realization of the microphones configuration that will be used throughout this chapter. For the source position, we will place it randomly in the room. We will also stick with a shoe box shaped room as the geometry that we will estimate.

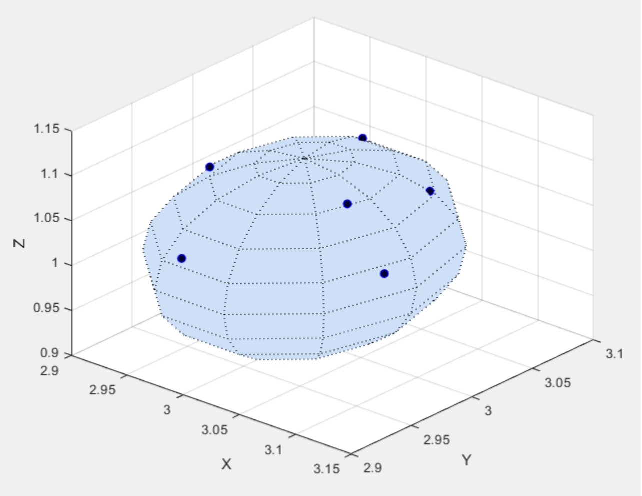


Figure 4.1: A realization of microphones configuration on the surface of a sphere with radius 0.10 m

For clarity purpose, there are important pre-conditions that we assume:

1. Availability of the room impulse response.

2. The microphones position is known a priori.

3. The sources are not collocated with the microphones.

4. The influence of the head and shoulders of a person or a robot on the RIR is ignored.

5. There is no occlusion between the sources and microphones.

Based on these points, we did the simulation in this chapter. The next section will introduce the problem that emerges due to the closed microphones configuration and its solution.

## 4.2 A Problem due to The Sphere's surface Microphones Configuration

Both the graph based method[6] and the greedy subspace method[7] depends on the **R** (microphones position) matrix in their source position localization technique. In [6], the adopted Pollefeys method uses rank 5 approximation (in 3D) on the squared distance estimation matrix (**D**) while the least square method in [7] multiplies the pseudoinverse of the **R** matrix to the **D** matrix. These reliances indicate that if the **R** matrix is not invertible, it will affect the accuracy of the sources' position estimation result ($\hat{\mathbf{S}}$).

Unlike the previous chapter where microphones are randomly placed inside a room, a close microphones arrangement increases the microphones position interdependence and reduces the information that is contained in the **R** matrix. Basically, it will make all singular values of **R** decrease. The largest decrement is experienced by the smallest singular values of **R**. As a result, the rank of the **R** matrix will reduce. In a three-dimensional case where the microphones are put on the surface of a sphere, the rank of the **R** matrix and the squared EDM of the inter-microphones distance becomes 4. This fact introduces a low rank dimension problem which is especially harmful to the image source localization method because both Pollefeys and the least squares depend on the **R** matrix.

The Pollefeys method is based on the rank 5 approximation, the rank deficiency in the **R** matrix causes a failure in the Cholesky factorization step of the Pollefeys method (see Appendix A) which leads to a wrong estimation of the source position matrix ($\hat{\mathbf{S}}$). For the least squares, since the 5th singular value of **R** matrix is almost zero, its inverse will blow up the solution and give incorrect ($\hat{\mathbf{S}}$) matrix. Although we face the same problem in both methods, we choose to resolve the problem in the least squares because the result in Chapter 3 clearly shows that the least squares has better accuracy than the Pollefeys.

## 4.3 Improving The Least Squares as The Image Sources Localization Technique

The original least square method results in a false estimation of the image sources position because of non invertible $\mathbf{R}$ matrix. Recall the structure of $\mathbf{R}$ and $\mathbf{S}$ matrix,

$$\mathbf{R}^T = \begin{bmatrix} \mathbf{r}_1^T\mathbf{r}_1 & -2x_1 & -2y_1 & -2z_1 & 1 \\ \mathbf{r}_2^T\mathbf{r}_2 & -2x_2 & -2y_2 & -2z_2 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{r}_m^T\mathbf{r}_m & -2x_m & -2y_m & -2z_m & 1 \end{bmatrix} \in \mathbb{R}^{M\times 5}, \tag{4.1}$$

$$\mathbf{S} = \begin{bmatrix} 1 & 1 & \cdots & \cdots & 1 \\ x_1 & x_2 & \cdots & \cdots & x_n \\ y_1 & y_2 & \cdots & \cdots & y_n \\ z_1 & z_2 & \cdots & \cdots & z_n \\ \mathbf{s}_1^T\mathbf{s}_1 & \mathbf{s}_2^T\mathbf{s}_2 & \cdots & \cdots & \mathbf{s}_n^T\mathbf{s}_n \end{bmatrix} \in \mathbb{R}^{5\times N}, \tag{4.2}$$

and their relation with the squared distance data matrix ($\mathbf{D}$):

$$\mathbf{R}^T\mathbf{S} = \mathbf{D} \ \in \mathbb{R}^{M\times N}. \tag{4.3}$$

Thanks to the special structure that the matrix $\mathbf{S}$ (Eq.4.2) has, its first row is the all ones vector ($\mathbf{1} \in \mathbb{R}^{1\times N}$), we can modify the original least square problem to the equality constraint least square problem by transforming that special structure into an equality constraint. The equality constraint forces the first row of our estimated $\mathbf{S}$ matrix ($\hat{\mathbf{S}}$) to be the all ones vector ($[1\ ,1\ ,\cdots ,1]^T$). The new formulated problem is given by

$$\begin{aligned} &\underset{\hat{\mathbf{k}}}{\text{minimize}} \quad \|\mathbf{R}^T\hat{\mathbf{k}} - \hat{\mathbf{d}}\|_2^2 \\ &\text{subject to} \quad \mathbf{a}^T\hat{\mathbf{k}} = 1, \end{aligned} \tag{4.4}$$

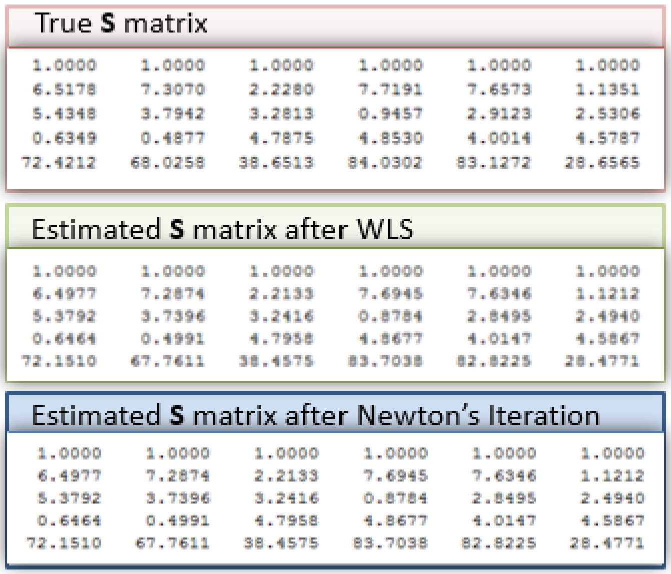
where $\mathbf{R}^T$ is the microphones position matrix, $\hat{\mathbf{k}}$ is a vector that we want to estimate, $\hat{\mathbf{k}} = [1\ \hat{x}_1\ \hat{y}_1\ \hat{z}_1\ \hat{\mathbf{s}}_n^T\hat{\mathbf{s}}_n]^T$ , $\hat{\mathbf{d}}$ is a column from the estimated true echoes combination matrix, $\hat{\mathbf{D}}$ (the output of the echo labeling solver block), and $\mathbf{a}$ is the $[1\ 0\ 0\ 0\ 0]^T$ vector.

The minimization problem in Eq. 4.4 is a convex problem with an equality constraint so it can be solved by applying Newton's iteration as described in [17]. Although the Newton iteration can solve the problem, the estimation result depends on the selected starting point. If the chosen starting point is poor, the outcome might be imprecise. To offer a good initial point for Newton's iteration, we use an approximate solution of the problem in Eq.4.4 based on the unconstrained least squares problem:

$$\underset{\hat{\mathbf{k}}}{\text{minimize}} \left\| \begin{pmatrix} \mathbf{R}^T \\ \omega\mathbf{a} \end{pmatrix} \hat{\mathbf{k}} - \begin{pmatrix} \hat{\mathbf{d}} \\ \omega 1 \end{pmatrix} \right\|_2^2 , \tag{4.5}$$

for large weight $\omega$ ($\omega \gg 1$). This problem is equal to the weighted least squares (WLS) problem [18],

$$\underset{\hat{\mathbf{k}}}{\text{minimize}} \left\| \mathbf{D}_\omega(\mathbf{F}\hat{\mathbf{k}} - \mathbf{g}) \right\|_2^2 , \tag{4.6}$$



True **S** matrix

```
 1.0000    1.0000    1.0000    1.0000    1.0000    1.0000
 6.5178    7.3070    2.2280    7.7191    7.6573    1.1351
 5.4348    3.7942    3.2813    0.9457    2.9123    2.5306
 0.6349    0.4877    4.7875    4.8530    4.0014    4.5787
72.4212   68.0258   38.6513   84.0302   83.1272   28.6565
```

Estimated **S** matrix after WLS

```
 1.0000    1.0000    1.0000    1.0000    1.0000    1.0000
 6.4977    7.2874    2.2133    7.6945    7.6346    1.1212
 5.3792    3.7396    3.2416    0.8784    2.8495    2.4940
 0.6464    0.4991    4.7958    4.8677    4.0147    4.5867
72.1510   67.7611   38.4575   83.7038   82.8225   28.4771
```

Estimated **S** matrix after Newton's Iteration

```
 1.0000    1.0000    1.0000    1.0000    1.0000    1.0000
 6.4977    7.2874    2.2133    7.6945    7.6346    1.1212
 5.3792    3.7396    3.2416    0.8784    2.8495    2.4940
 0.6464    0.4991    4.7958    4.8677    4.0147    4.5867
72.1510   67.7611   38.4575   83.7038   82.8225   28.4771
```

Figure 4.2: The true $S$ matrix (the first square), its estimation result using WLS (the second square), and the Newton iteration (the third square).

where

$$\mathbf{F} = \begin{pmatrix} \hat{\mathbf{R}}^T \\ \mathbf{a} \end{pmatrix} \in \mathbb{R}^{(M+p)\times 5}, \quad \mathbf{g} = \begin{pmatrix} \hat{\mathbf{d}} \\ 1 \end{pmatrix} \in \mathbb{R}^{(M+p)},$$

and

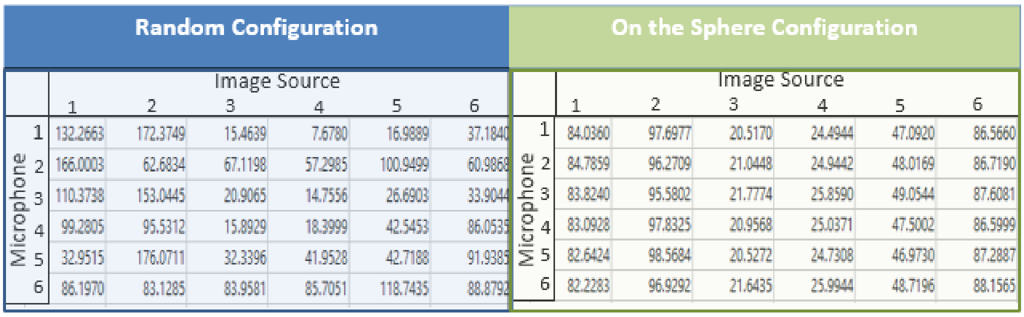
$$\mathbf{D}_\omega = diag(\underbrace{1, \cdots, 1}_{M}, \underbrace{\omega, \cdots, \omega}_{p}).$$

We just have one linear constraint so $p$ is equal to one, but in general $p \geq 1$.

When the output of Eq.4.5 serves as an initial point for the Newton iteration, the point does not change anymore (Figure 4.2). In this case we set $\omega = 1000$. This means the approximated solution from the WLS already gave the optimum solution. These results reveal that the WLS can be used as an alternative for the Newton iteration to localize the image sources position and fix our low dimension problem that is caused by the new microphones configuration. In the rest of this chapter, the WLS will be utilized as the image sources position localization.

## 4.4 Variance Filter

The new microphone configuration on the sphere's surface raises the low dimension problem which affects the image source localization method. On the other hand, it also brings an advantage that is useful in reducing the computational cost for estimating the room geometry. Unlike in the random microphones configuration where the squared distances between a particular image source to the microphones vary a lot,



**Random Configuration**

| | Image Source 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Microphone 1 | 132.2663 | 172.3749 | 15.4639 | 7.6780 | 16.9889 | 37.1840 |
| 2 | 166.0003 | 62.6834 | 67.1198 | 57.2985 | 100.9499 | 60.9868 |
| 3 | 110.3738 | 153.0445 | 20.9065 | 14.7556 | 26.6903 | 33.9044 |
| 4 | 99.2805 | 95.5312 | 15.8929 | 18.3999 | 42.5453 | 86.0535 |
| 5 | 32.9515 | 176.0711 | 32.3396 | 41.9528 | 42.7188 | 91.9385 |
| 6 | 86.1970 | 83.1285 | 83.9581 | 85.7051 | 118.7435 | 88.8792 |

**On the Sphere Configuration**

| | Image Source 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Microphone 1 | 84.0360 | 97.6977 | 20.5170 | 24.4944 | 47.0920 | 86.5660 |
| 2 | 84.7859 | 96.2709 | 21.0448 | 24.9442 | 48.0169 | 86.7190 |
| 3 | 83.8240 | 95.5802 | 21.7774 | 25.8590 | 49.0544 | 87.6081 |
| 4 | 83.0928 | 97.8325 | 20.9568 | 25.0371 | 47.5002 | 86.5999 |
| 5 | 82.6424 | 98.5684 | 20.5272 | 24.7308 | 46.9730 | 87.2887 |
| 6 | 82.2283 | 96.9292 | 21.6435 | 25.9944 | 48.7196 | 88.1565 |

Figure 4.3: The squared distance (**D**) matrix between the microphones and the image sources for the random and the sphere's surface microphones configuration when $M = 6$, $N = 6$, and room size = [8 6 5]m. The position of the true source is the same for both microphones configurations.

the sphere's surface microphones configuration with small radius shrinks this variability of the squared distance. In other words, the new microphones constellation reduce the variance between elements in each column of the squared distance matrix (**D**). The squared distance matrix (**D**) for both random and sphere's surface microphones configuration with radius 10 cm are depicted in Fig.4.3. Note that, for the random microphones configuration, the difference between elements in a column of the **D** matrix is very large while for the sphere's surface microphones configuration the difference is very small.

Using the above observation, we can determine roughly which columns in the candidate of the squared distance combination, $\tilde{\mathbf{D}}$ that do not belong to the true echoes combination beforehand. In this way, we can equip the echo labeling solver step with an initial step called variance filter as illustrated in the second box of Fig.4.4. The implementation of this variance filter is by introducing a variance threshold called $v$, the output of this filter is all columns of $\tilde{\mathbf{D}}$ that have variance smaller than $v$ called $\tilde{\mathbf{D}}_v$. The size of $v$ depends on the radius of the sphere. The larger the radius, the larger $v$ will be.

## 4.5 Implementation

Recall from Chapter 3 about the echo labeling solver, Figure 4.5 displays all possible techniques (original and combination) provided by Jager's and Coutino's methods for estimating the room geometry. In the previous chapter, we already saw their performance for the random microphones configuration. Now, holding the pre-conditions that were mentioned in Section 4.1 we will test their performance for the new microphones configuration on the sphere's surface.

The room that will be used for this simulation is a shoe box shaped room with dimension $8m \times 6m \times 5m$. The performance of the five methods are assessed based on the vertex estimation result and the computational cost of 100 experiments with the following set up:

1. The number of sources ($N$) is varied, $M = 6$, $\sigma = 0.001m$, sphere radius ($r =$

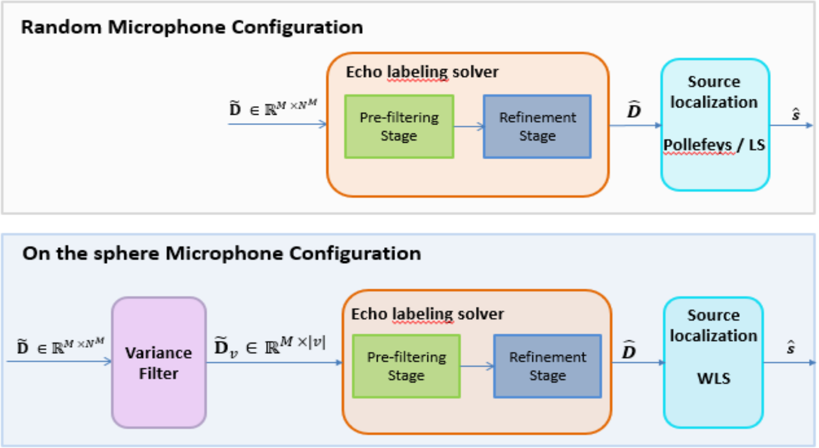

Random Microphone Configuration
$\tilde{\mathbf{D}} \in \mathbb{R}^{M \times N^M}$
Echo labeling solver
Pre-filtering Stage
Refinement Stage
$\widehat{\boldsymbol{D}}$
Source localization
Pollefeys / LS
$\hat{\boldsymbol{s}}$
On the sphere Microphone Configuration
$\tilde{\mathbf{D}} \in \mathbb{R}^{M \times N^M}$
Variance Filter
$\tilde{\mathbf{D}}_v \in \mathbb{R}^{M \times |v|}$
Echo labeling solver
Pre-filtering Stage
Refinement Stage
$\widehat{\boldsymbol{D}}$
Source localization
WLS
$\hat{\boldsymbol{s}}$

Figure 4.4: The block diagram of estimating the image source position for the random microphones configuration (top box) and the sphere's surface microphones configuration (bottom box)

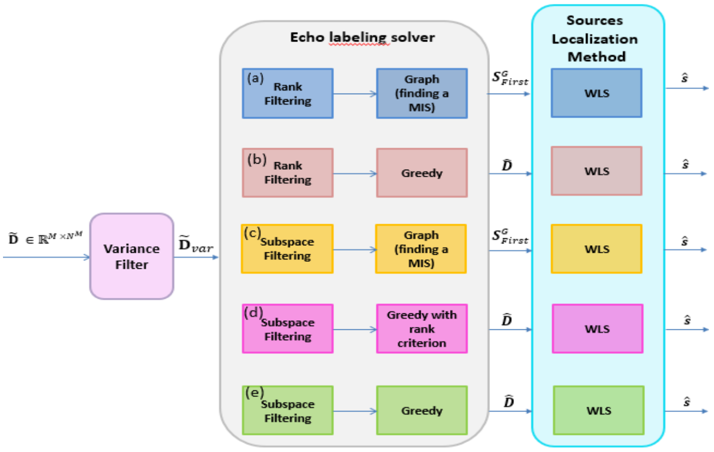


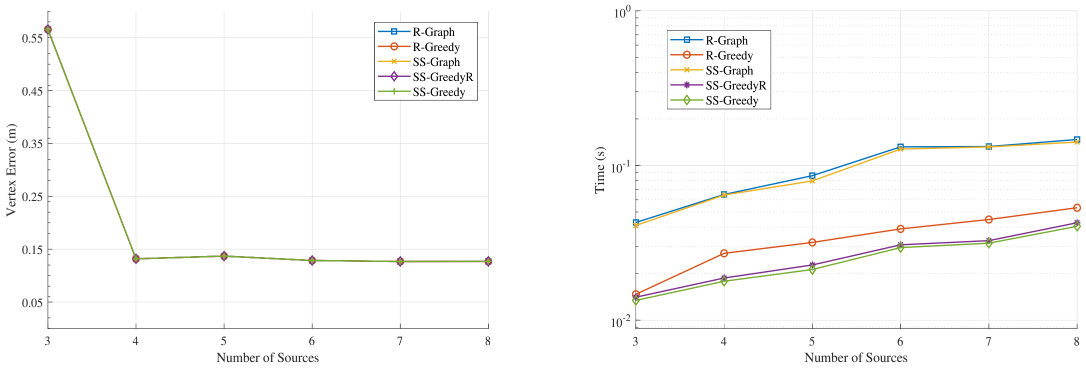
Figure 4.5: All possible combinations of the echo labeling solver that will be implemented for estimating the room geometry with the sphere's surface microphones configuration.

$10cm$).

2. The number of microphones ($M$)is varied, $N = 6$, $\sigma = 0.001m$, sphere radius ($r = 10cm$).

3. The radius of the sphere ($r$) is varied, $M = 6$, $N = 6$, $\sigma = 0.001m$.

Close distance microphones arrangement makes the echo labeling solver algorithms more sensitive towards the uncertainty in the distance data ($\sigma$). A small increment in the noise standard deviation ($\sigma$) affects the squared distance data and leads the algorithm to pick the wrong columns of $\tilde{\mathbf{D}}$. This happens especially when the sphere radius is smaller than 10 cm. Hence, in this experiment, we will fix the ($\sigma$) to 0.001m.



(a) Vertex estimation error VS number of sources (b) Computational Time VS number of sources

Figure 4.6: The number of sources varied

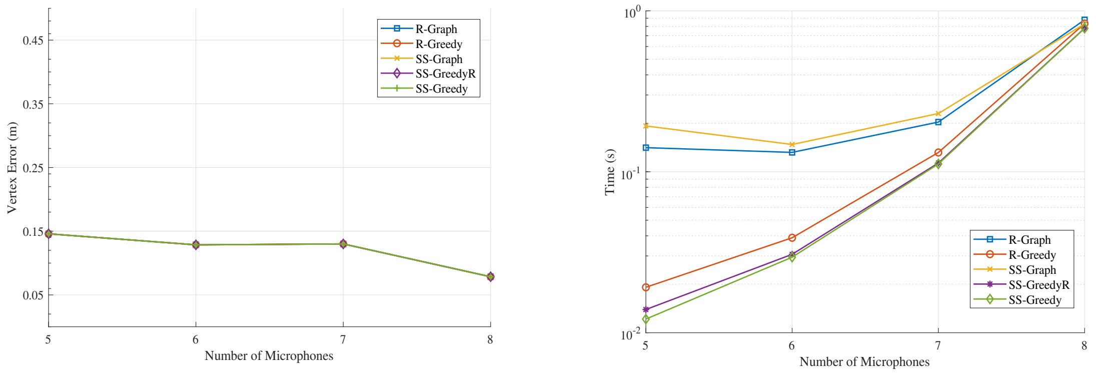
### 4.5.1 The Number of Sources

Figure 4.6 depicts the performance of the five methods in terms of accuracy (Fig. 4.6a) and computational time (Fig. 4.6b). The estimation accuracy improves as the number of sources increases but the improvement is insignificant after $N > 4$. The abrupt decrement of vertex estimation error takes place when $N$ increases from 3 to 4. This situation takes place because the vertex estimation method needs at least two wall points from two sources. When $N = 3$, if the image sources estimation result from two sources are bad, we are left with one valid wall point from one source. Thus, the vertex estimation method gives a wrong result. If $N = 4$ we have more wall points that can produce better vertex estimation result. The increment in the computational time is linear with the number of sources due to the time needed for running the algorithm for each source.

The five algorithms for solving echo labeling problem (Fig.4.5) have a similar achievement in terms of accuracy as illustrated in Fig.4.6a. On the contrary, not all combination methods have the same computational time. Fig. 4.6b shows that the lowest computational time was held by the 5th algorithm (e) while the highest was the 3rd algorithm (c). We can infer that the algorithms which contain the graph method as their refinement step have higher computational time than the algorithms with the greedy procedure. This condition also happens for the algorithms that use rank filtering as their pre-filtering step. These behaviors are in line with the result of Table 3.1.

### 4.5.2 The Number of Microphones

Figure 4.7a illustrates that the vertex estimation error decreases as the number of microphones increases. Intuitively, adding the number of microphones will improve the vertex estimation error. unfortunately, due to the exponential increment in the number of columns in the squared distance matrix $(\tilde{\mathbf{D}})$, we face a memory limitation problem in Matlab to perform the simulation. We can overcome this problem by dividing the



(a) Vertex estimation error VS Number of microphones

(b) Computational Time VS number of microphones

Figure 4.7: The number of microphones varied

$\tilde{\mathbf{D}}$ matrix into smaller matrices with fewer columns, then pass it to the next algorithm block by block. However, this process still consumes a lot of time. To prove our first guess about the behavior of the vertex estimation curve when $M$ increase, we did a simulation which bypasses the echo labeling solver block and assumed that we have the correct echoes combination (the $\mathbf{D}$ matrix), then directly use it to estimate the room geometry. The simulation result is displayed in Fig. 4.8. This curve does not only support our intuition but also provides another fact that the estimation error does not have much improvement after $M = 8$.

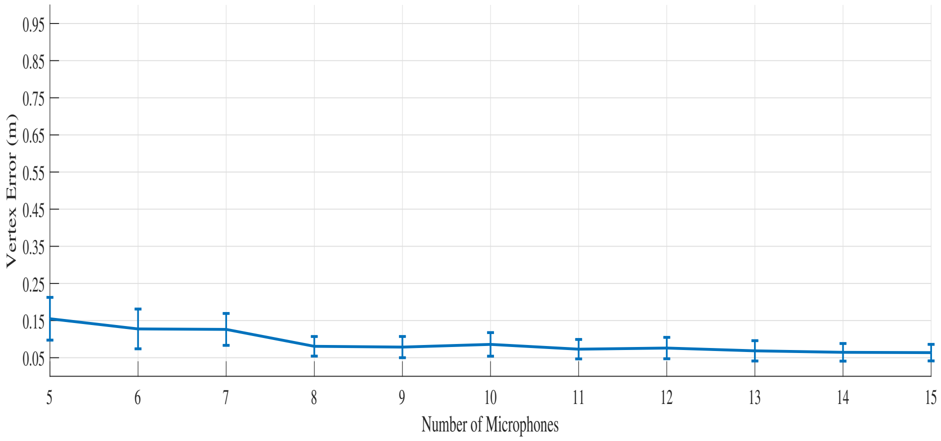
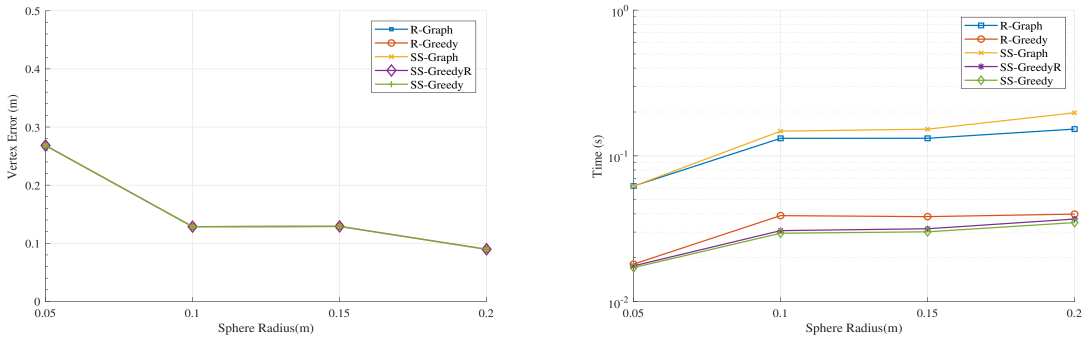


Figure 4.8: The vertex estimation error VS the number of microphones when $N = 6$, $\sigma = 0.001m$, $r = 10cm$ with the apriori knowledge of the squared distance data matrix.

The vertex estimation error for the five algorithms is the same while the computational time curve gives a steep increase especially for algorithm b,c,e. However, as the number of microphones reaches 8, the computational time curve for all algorithms are met because the number of columns in $\tilde{\mathbf{D}}$ is very large (in this case $6^8$ columns). Therefore, most of the time is consumed in the variance filter step and the time for other steps become negligible.

### 4.5.3 The Radius of The Sphere

Intuitively, as the radius of the sphere becomes large, the inter-dependency between the microphones decreases and the condition of $R$ matrix is better (the smallest singular value climb up so the $\mathbf{R}$ matrix is close to a full rank matrix). As a result, the vertex estimation error declines(Fig.4.9a). For the computational time, it grows slowly (Fig. 4.9b). The large increment takes place when the radius goes up from 5cm to 10 cm. This occurs because the initial threshold for the variance filter ($v$) and the rank filter ($\epsilon$) at r = 5cm are quite small. When $r$ changes to 10cm, the previous initial threshold becomes too small and cannot produce the expected result, so the algorithm adaptively increases the threshold until it reaches the expected result and costs more time to complete.



(a) Vertex estimation error VS sphere radius

(b) Computational Time VS sphere radius

Figure 4.9: The sphere radius varied

## 4.6 Summary

From the experiments that have been done, some important points can be inferred:

- The WLS solves the low dimension problem in the sphere's surface microphones constellation.
- The five algorithms in Fig.4.5 for solving the echo labeling problem share the same performance in terms of the vertex estimation error.
- In terms of the computational cost, the fastest algorithm is held by algorithms (d) and (e) of Fig.4.5.

- For a sphere with radius 10 cm, the smallest error that can be achieved is 6 cm with 8 microphones and computational time of 1 second.

# Conclusions 5

The work presented in this thesis focused on improving the two current methods; the graph-based and the subspace-greedy method for estimating the room geometry. We also construct hybrid methods by interchanging the intermediate step of the two methods. Moreover, we verify the feasibility of the hybrid methods for: (i) the random microphones configuration and (ii) the sphere's surface microphones configuration with a small radius. The performance of all methods is assessed through their accuracy and computational cost.

In Chapter 3, we modify the graph based method by replacing its source localization step from Pollefeys to the least squares, then we substitute the graph procedure of finding a maximum independent set with the greedy procedure in the refinement step. In this way, we end up with the following combination methods:

a. rank - graph - Pollefeys (the original graph based method)

b. rank - graph - LS

c. rank - greedy - Pollefeys

d. rank - greedy - LS

The experimental results with the random microphones configuration showed that the least squares outperform the Pollefeys method for localizing image sources position which leads to lower vertex estimation error and computational cost. The least squares also alleviate the minimum source requirement of the Pollefeys method. Furthermore, the fourth method (d) attained comparable accuracy with respect to the second method (b) but restrict the minimum number of microphones to 6 ($M \geq 6$).

In Chapter 4, we fully implemented the hybrid techniques which are derived from the graph and the greedy subspace method for estimating the room geometry with microphones located on the sphere's surface. The new microphones constellation rises a low dimensional problem which caused the output of the source localization step inaccurate. To handle this problem we substitute the original least squares method with the equality constrained least squares (approximated by the weighted least squares). Moreover, the insertion of the variance filter as the preliminary step of the echo labeling solver attains a great reduction in the total computational time for all techniques. The sequence of the derived techniques are:

a. variance - rank - graph - weighted least squares

b. variance - rank - greedy - weighted least squares

c. variance - subspace - graph - weighted least squares

d. variance - subspace - greedy(rank) - weighted least squares

e. variance - subspace - greedy - weighted least squares

Finally, the experimental results proved that all techniques (a-e) are feasible to estimate the vertex of the room with centimeter precision within seconds.

## 5.1 Future Directions

The following points are the ideas that can benefit further research on this topic:

- All the experiments in this thesis used the synthetic squared distance data, so we skipped the room impulse response acquisition procedure and the time of arrival (TOA) estimation step. In the future, for the sphere's surface microphone configuration, it will be nice to check the performance of the derived algorithms if the squared distance data are extracted from the room impulse response since the TOA estimation procedure plays an important role on the final accuracy of the room geometry estimation.

- Since the final goal of this research is to check the feasibility of the available methods for gaming application and robot navigation, considering the presence of objects and people inside the room are important matters that can provide more details in the room geometry reconstruction.

- Of further interest is also including the head transfer function for placing the microphones on the helmet because, in the real situation, the reflection from the head and shoulder of a person or a robot affects the signal that is received by the microphone. Moreover, since the people or a robot will always move, leaving the assumption of knowing the microphones' position is preferable. Thus, instead of the microphones' position, we just consider the pairwise distance of the microphones.

- Our approach in estimating the room geometry is considered as an image-source reversion method. In the future, it would be beneficial if we can compare our approach with a direct reflector localization method as proposed in [9].

# Pollefeys Method for Sources and Microphones Localization A

This appendix contains a brief explanation of the adapted Pollefeys method which is used in the graph based method [6]. For the complete derivation of this method, the readers are referred to [15].

## A.1 Preliminary Equations

Consider $M$ microphones (indexed by $m$) and $N$ sources (indexed by $n$) with their spatial coordinates defined by $\mathbf{r}_m = [x_m, y_m, z_m]^T \in \mathbb{R}^3$ and $\mathbf{s}_n = [X_n, Y_n, Z_n]^T \in \mathbb{R}^3$ respectively. The squared distance between the $(m, n)$-th microphone and source pair can be formulated as :

$$d_{m,n} = (x_m - X_n)^2 + (y_m - Y_n)^2 + (z_m - Z_n)^2 \, . \tag{A.1}$$

Expressed in vector notation, Eq. A.1 becomes

$$d_{m,n} = \mathbf{R}_m^T \mathbf{S}_n \, , \tag{A.2}$$

where

$$\begin{aligned} \mathbf{R}_m &= [\mathbf{r}_m^T \mathbf{r}_m \;\; -2x_m \;\; -2y_m \;\; -2z_m \;\; 1]^T \in \mathbb{R}^{5\times 1} \, , & \text{(A.3)} \\ \mathbf{S}_n &= [1 \; X_n \; Y_n \; Z_n \; \mathbf{s}_n^T \mathbf{s}_n] \in \mathbb{R}^{5\times 1} \, . & \text{(A.4)} \end{aligned}$$

Stacking up all the $(m, n)$ pair squared distance $d_{m,n}$ yields a squared distance matrix $\mathbf{D} \in \mathbb{R}^{M\times N}$ and Eq. A.2 can be expressed in a matrix form as

$$\mathbf{R}^T \mathbf{S} = \mathbf{D} \in \mathbb{R}^{M\times N} \, , \tag{A.5}$$

where $\mathbf{R} = [\mathbf{R}_1, ..., \mathbf{R}_M]$ and $\mathbf{S} = [\mathbf{S}_1, ..., \mathbf{S}_N]$ are the microphone and source position matrices according to Eq. A.3 and A.4.

## A.2 Sources and Microphones Localization

Pollefeys method used rank 5 approximation of the $\mathbf{D}$ matrix in Eq.A.5 to recover the $\mathbf{R}$ and $\mathbf{S}$ matrices. The approximation is started by computing the singular value decomposition (SVD) of $\mathbf{D}$,

$$\mathbf{D} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T \, . \tag{A.6}$$

Defining $\hat{\mathbf{R}} = \mathbf{U}^T$ and $\hat{\mathbf{S}} = \boldsymbol{\Sigma}\mathbf{V}^T$, we have

$$\hat{\mathbf{R}}^T \hat{\mathbf{S}} = \mathbf{D}, \tag{A.7}$$

where $\hat{\mathbf{R}}$ and $\hat{\mathbf{S}}$ are related to $\mathbf{R}$ and $\mathbf{S}$ by a transformation matrix $\mathbf{H}$,

$$\mathbf{R}^T\mathbf{S} = \hat{\mathbf{R}}\mathbf{H}^{-1}\mathbf{H}\hat{\mathbf{S}}\,. \tag{A.8}$$

The transformation matrix $\mathbf{H}$ is consisted of three concatenate transformation matrices, i.e.,

$$\mathbf{H} = \mathbf{H}_Q\mathbf{H}_R\mathbf{H}_S. \tag{A.9}$$

Each transformation matrix corresponds to preserve the special structures that are possessed by $\mathbf{R}$ and $\mathbf{S}$ matrices. The first transformation, $\mathbf{H}_S$ will ensure that the first row of $\mathbf{S}$ is all ones vector. $\mathbf{H}_S$ can be written as

$$\mathbf{H}_S = \begin{bmatrix} \mathbf{h}_S^T \\ 0 \quad \mathbf{I} \end{bmatrix}, \tag{A.10}$$

$\mathbf{h}_S^T$ can be found by solving the linear system of equations $\mathbf{h}_S^T\hat{\mathbf{S}} = \mathbf{1}$. this step requires at least 5 microphones. In the same manner, the second transformation $\mathbf{H}_R$ preserves the all ones vector structure in the last row of $\mathbf{R}$. $\mathbf{H}_R$ can be formulated as

$$\mathbf{H}_R = \begin{bmatrix} \mathbf{I} & \\ 0 & \mathbf{h}_R^T \end{bmatrix}^{-1}. \tag{A.11}$$

By solving $\hat{\mathbf{R}}^T\mathbf{H}_S^{-1}\mathbf{h}_R = \mathbf{1}$ $\mathbf{h}_R$ can be computed. The last transformation $\mathbf{H}_Q$ imposes the quadratic consistency constraints on $\mathbf{R}$ and $\mathbf{S}$. We can derive the constrain on $\mathbf{S}_n$ as

$$\mathbf{S}_n^T\,\mathbf{B}\,\mathbf{S}_n = 0 \text{ with } \mathbf{B} = \begin{bmatrix} 0 & 0 & 0 & 0 & -\frac{1}{2} \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ -\frac{1}{2} & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{A.12}$$

Therefore,

$$\hat{\mathbf{S}}_n^T\,\mathbf{H}^T\,\mathbf{B}\,\mathbf{H}\,\hat{\mathbf{S}}_m = 0\,. \tag{A.13}$$

Now, we define $\hat{\mathbf{S}}_n^{'} = \mathbf{H}_R\,\mathbf{H}_S\,\hat{\mathbf{S}}_n$ and $\mathbf{Q} = \mathbf{H}_Q^T\,\mathbf{B}\,\mathbf{H}_Q$ to obtain the following linear equation:

$$\hat{\mathbf{S}}_n^{'}\,\mathbf{Q}\,\hat{\mathbf{S}}_n^{'} = 0\,, \tag{A.14}$$

for determining the coefficients of the symmetric matrix $\mathbf{Q}$. Since we have to keep the first row of $\hat{\mathbf{S}}_n$ and the last row of $\hat{\mathbf{R}}_m$ unchanged, the transformation $\mathbf{H}_Q$ must have the following form:

$$\mathbf{H}_Q = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ . & . & . & . & 0 \\ . & . & . & . & 0 \\ . & . & . & . & 0 \\ . & . & . & . & 1 \end{bmatrix} \text{ and } \mathbf{Q} = \begin{bmatrix} . & . & . & . & -\frac{1}{2} \\ . & . & . & . & 0 \\ . & . & . & . & 0 \\ . & . & . & . & 0 \\ -\frac{1}{2} & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{A.15}$$

Considering that $\mathbf{Q}$ is symmetric, it has ten degrees of freedom. These ten coefficients can be computed linearly with at least ten sources using Eq.A.14. Then, $\mathbf{H}_Q$ can be formed using the entries of $\mathbf{Q}$,

$$\mathbf{H}_Q = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & & & & 0 \\ 0 & & \mathbf{K} & & 0 \\ 0 & & & & 0 \\ -Q_{11} & -2Q_{12} & -2Q_{13} & -2Q_{14} & 1 \end{bmatrix} , \tag{A.16}$$

where $\mathbf{K}$ is the Cholesky factorization of the middle part ($3 \times 3$) of $\mathbf{Q}$. Since all transformations have been derived, we can find:

$$\mathbf{R} = (\hat{\mathbf{R}}^T \, \mathbf{H}_S^{-1} \, \mathbf{H}_R^{-1} \, \mathbf{H}_Q^{-1})^T , \tag{A.17}$$

$$\mathbf{S} = \mathbf{H}_Q \, \mathbf{H}_R \, \mathbf{H}_S \, \hat{\mathbf{S}} , \tag{A.18}$$

from which the position of the microphones and the sources can be extracted.

# Bibliography

[1] Mario Alberto Coutino. *Identification of Room Boundaries for Sound Field Estimation*. PhD thesis, TU Delft, 2016.

[2] D. N. Zotkin, R. Duraiswami, and L. S. Davis. Rendering localized spatial audio in a virtual auditory space. *IEEE Transactions on Multimedia*, 6(4):553–564, Aug 2004. ISSN 1520-9210. doi: 10.1109/TMM.2004.827516.

[3] Yiteng Huang, J. Benesty, G. W. Elko, and R. M. Mersereati. Real-time passive source localization: a practical linear-correction least-squares approach. *IEEE Transactions on Speech and Audio Processing*, 9(8):943–956, Nov 2001. ISSN 1063-6676. doi: 10.1109/89.966097.

[4] Jont B. Allen and David A. Berkley. Image method for efficiently simulating small room acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950, April 1979. doi: 10.1121/1.382599.

[5] I. Dokmanic, R. Parhizkar, A. Walther, Y. M. Lu, and M. Vetterli. Acoustic echoes reveal room shape. *Proceedings of the National Academy of Sciences*, 110 (30):12186–12191, 2013. doi: 10.1073/pnas.1221464110.

[6] I. Jager, R. Heusdens, and N. D. Gaubitch. Room geometry estimation from acoustic echoes using graph-based echo labeling. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, March 2016. doi: 10.1109/ICASSP.2016.7471625.

[7] M. Coutino, M. B. Mller, J. K. Nielsen, and R. Heusdens. Greedy alternative for room geometry estimation from acoustic echoes: A subspace-based method. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 366–370, March 2017. doi: 10.1109/ICASSP.2017.7952179.

[8] Heinrich Kuttruff. *Room acoustics*. CRC Press, 6 edition, 2017.

[9] L. Remaggi, P. J. B. Jackson, P. Coleman, and W. Wang. Acoustic reflector localization: Novel image source reversion and direct localization methods. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(2):296–309, Feb 2017. ISSN 2329-9290. doi: 10.1109/TASLP.2016.2633802.

[10] Ivan Dokmanic, Reza Parhizkar, Juri Ranieri, and Martin Vetterli. Euclidean distance matrices. *IEEE Signal Processing Magazine*, page 1230, Oct 2015.

[11] Ivan Dokmanic, Yue M. Lu, and Martin Vetterli. Can one hear the shape of a room: The 2-d polygonal case. *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011. doi: 10.1109/icassp.2011.5946405.

[12] Fabio Antonacci, Jason Filos, Mark R. P. Thomas, Emanul A. P. Habets, Augusto Sarti, Patrick A. Naylor, and Stefano Tubaro. Inference of room geometry from

acoustic impulse responses. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(10):26832695, 2012. doi: 10.1109/tasl.2012.2210877.

[13] Luca Remaggi, Philip J. B. Jackson, Philip Coleman, and Wenwu Wang. Room boundary estimation from acoustic room impulse responses. *2014 Sensor Signal Processing for Defence (SSPD)*, 2014. doi: 10.1109/sspd.2014.6943328.

[14] Flvio Ribeiro, Dinei Florencio, Demba Ba, and Cha Zhang. Geometrically constrained room modeling with compact microphone arrays. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(5):14491460, 2012. doi: 10.1109/tasl.2011.2180897.

[15] M. Pollefeys and D. Nister. Direct computation of sound and microphone locations from time-difference-of-arrival data. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2445–2448, March 2008. doi: 10.1109/ICASSP.2008.4518142.

[16] John C Gower and Garmt B Dijksterhuis. Procrustes problems. 2004. doi: 10.1093/acprof:oso/9780198510581.001.0001.

[17] Stephen Boyd and Lieven Vandenberghe. *Convex optimization.* Cambridge University Press, 2015.

[18] Aleksandr Ivanovich Zhdanov and Sofya Yuryevna Gogoleva. Solving least squares problem with equality constraints based on augmented regularized normal equations. URL http://www.math.nthu.edu.tw/~amen/.