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

CAS-2016-4410475

M.Sc. Thesis

Identification of Room Boundaries for Sound Field Estimation

Mario Alberto Coutiño Minguez

Abstract

Echoes generated by the sound reflected off the walls of a room carry information about the geometry of the enclosure. Capitalization of this acoustic property could lead to improvements in current state-of-the-art methods for sound field estimation, where prior information can be used to improve the conditioning of the problem. In this thesis, robust and computational efficient methods are developed for identifying first order reflections to estimate the room geometry using small microphone arrays. Furthermore, as the estimation of such reflections becomes even more challenging in actual audio reproduction systems, this work aims to develop methods capable to deal with complications that might arise due to the employed drivers. This is done by considering the estimation problem in two different scenarios. Firstly, the first order reflections estimation problem is posed as a sorting problem. For this case, a set of echoes, received at different microphones, must be grouped accordingly to the wall which originated them. This problem is solved by using a greedy subspace-based algorithm. The proposed approach provides similar performance compared with the state-of-the-art method at a reduced computational cost. For the second scenario, instead of echoes, only raw microphones measurements are available. This instance of the problem is posed under an estimation theory framework, and solved by sequential minimization of a non-linear cost function based on the propagation of waves. Experimental results, evaluated in simulated shoe-box shaped rooms, demonstrate the performance and applicability of the proposed methods for room geometry estimation.

TUDelft

Faculty of Electrical Engineering, Mathematics and Computer Science

Delft University of Technology

Identification of Room Boundaries for Sound Field Estimation

Thesis

submitted in partial fulfillment of the
requirements for the degree of

Master of Science

in

Electrical Engineering

by

Mario Alberto Coutiño Minguez
born in La Paz, Baja California Sur, México

This work was performed at:

Bang & Olufsen A/S
Department of Research and Development
Struer, Denmark

Delft University of Technology

Delft University of Technology
Department of
Electrical 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 **“Identification of Room Boundaries for Sound Field Estimation”** by **Mario Alberto Coutiño Minguez** in partial fulfillment of the requirements for the degree of **Master of Science**.

Dated: 26th August 2016

Chairman: ________________________
dr.ir. R. Heusdens, TU Delft

Advisor: ________________________
M.B. Møller, M.Sc., Bang & Olufsen

Committee Members: ________________________
prof.dr.ir. G. Leus, TU Delft

____
prof. dr. E. Eisemann, TU Delft

Abstract

Echoes generated by the sound reflected off the walls of a room carry information about the geometry of the enclosure. Capitalization of this acoustic property could lead to improvements in current state-of-the-art methods for sound field estimation, where prior information can be used to improve the conditioning of the problem. In this thesis, robust and computational efficient methods are developed for identifying first order reflections to estimate the room geometry using small microphone arrays. Furthermore, as the estimation of such reflections becomes even more challenging in actual audio reproduction systems, this work aims to develop methods capable to deal with complications that might arise due to the employed drivers. This is done by considering the estimation problem in two different scenarios. Firstly, the first order reflections estimation problem is posed as a sorting problem. For this case, a set of echoes, received at different microphones, must be grouped accordingly to the wall which originated them. This problem is solved by using a greedy subspace-based algorithm. The proposed approach provides similar performance compared with the state-of-the-art method at a reduced computational cost. For the second scenario, instead of echoes, only raw microphones measurements are available. This instance of the problem is posed under an estimation theory framework, and solved by sequential minimization of a non-linear cost function based on the propagation of waves. Experimental results, evaluated in simulated shoe-box shaped rooms, demonstrate the performance and applicability of the proposed methods for room geometry estimation.

Acknowledgments

Looking back people always wonder how they arrived to a particular moment in their life. Luckily, the past has all the answers. It is just question of connecting the dots. Thinking of these last two years, I can only be grateful towards the amazing people that has surrounded me. I arrived the Netherlands without a clue of what I was going to do in Delft. However, during my stay I met wonderful people who made those windy and rainy days great experiences. A big share of what I have achieved until now has been due to them, and for that, I will always be in their debts. Thanks.

This work would not have been possible without the education I received from all my professors in Delft, whose teaching increased my interest in signal processing. In particular, I want to express my appreciation to Richard and Geert. The trust and confidence of Richard, my Delft thesis supervisor, gave me the opportunity to conduct my thesis at Bang & Olufsen. In addition, his feedback and support through the project made my work way easier. On the other side, the mentoring I received from Geert the past summer helped me to gain confidence in my work and to be more prepare to carry out independent research. I appreciate what both have done for me. Thanks.

During my project, I spent most of my time in Struer, Denmark (where Bang & Olufsen is located). In this small town in west Jutland, I came to appreciate things I have never thought about. My whole gratitude goes to the town that hosted me, such a peaceful and lovely place.

Few words are not enough to describe the experience at Bang & Olufsen. The atmosphere, the work, the passion. It was wonderful to experience a company whose employees bleed the brand. My respect and gratitude to all the people in the company which made my stay such a nice experience. Particularly, I would like to extend my gratitude to both Martin and Jesper, my company supervisors. The sharp insights in my thesis, and their knowledge in signal processing and acoustic shaped the thesis I am presenting today. Thanks to them I have learned the specifics of acoustics and sound experience unknown before to me. In addition, I would like to thank Søren, head of the research department, for all the help with respect to my time in Bang & Olufsen.

It has to be said that the winter was hard on me. Tons of rain, wind and very short (and dark) days, but thanks to the rest of the interns in the company my days were never boring. Thanks for such nice days and great experiences. Particularly for those days in which we gathered to have diner and discuss anything that crossed our minds. Those were great times.

Finally, I want to say that this work is dedicated to my family and the important people that I have left behind in Mexico. They know how deeply grateful I am to them for everything they have done/given up for me. They always think highly of me, supporting me in every step. The result of my hard work is the least I can offer to repay all their kindness.

Mario Alberto Coutiño Minguez
Delft, The Netherlands, 26th August 2016

Når man føler hvor lidet
man nåer med sin flid,
er det nyttigt at mindes, at
Ting Tar Tid.
-Piet Hein

When you feel how depressingly
slowly you climb,
it's well to remember that
Things Take Time.
-Piet Hein

Contents

List of Figures

List of Tables

Introduction 1

In the last years, the way in which we experience sound has started to change. Interest in extending the spatial properties of the current sound systems has increased, and along the way, so has the difficulty in achieving such goals. For example, beyond the traditional stereo systems, currently there is the possibility of developing individual sound experiences through sound zones [5][23]. In order to achieve this, a set of loudspeakers, distributed in a room, are employed to control the sound field in different spatial locations. This provides the possibility of isolating acoustic events in space. An example of a sound zone setup is shown in Fig. 1.1. In this instance three different sound zones allow people to enjoy two different programs in the chair and sofa, without disturbing the on-going conversations at the dining table.

Establishing such acoustic zones is a challenging problem that requires knowledge of the enclosure where the sound is reproduced. Boundaries in the reproduction environment introduce reflections that create standing waves, especially at low frequencies. The knowledge of how these reflections occur is crucial to the control of the sound field at the sound zones. In order to describe how the enclosure contributes to establish a sound field, current approaches measure the room impulse response (RIR) at different spatial locations. The RIRs describe the propagation of sound from the source location to a point in the space. They summarize the interaction of the direct path and the reflections occurring in the enclosure. These RIRs are then used to define filters to control the sound field at the zones of interest. However, in practical situations this will imply that it would be necessary to measure every distinct room to set up a system. Furthermore, as the zones of interest are not constrained to a single point in space, a large quantity of measurements will be required to completely cover the zones of interest. This proves itself to be a costly task in both time and resources.

Therefore, it would be of interest, in an audio reproduction setup, to be able to esti-

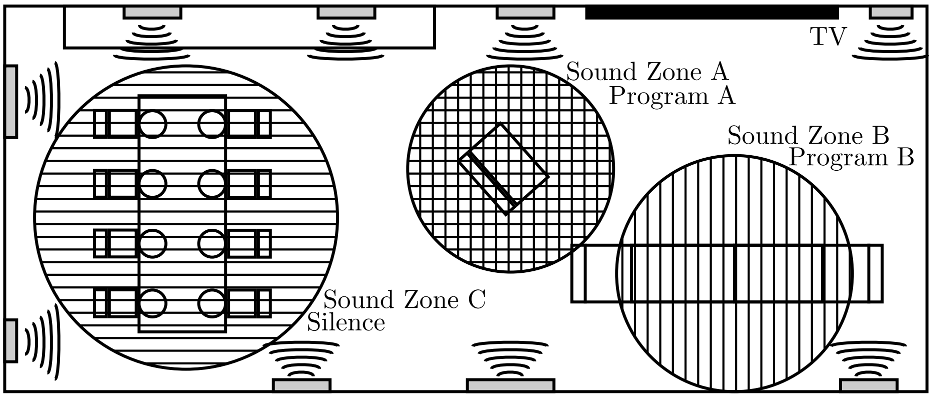


Figure 1.1: Configuration of individual sound zones in an arbitrary room

mate the sound field in a room given a set of measurements acquired using microphones embedded in the available loudspeakers. With this information the RIRs at any point in the enclosure could be estimated and then used to design the filters for controlling the sound field. This can be achieved considering that the sound field in a room is defined by the contributions of the direct path and the reflections due to its boundary. Hence, if we are able to estimate the shape of the reproduction environment, i.e., walls locations, it is possible to predict how the sound field is at an arbitrary point inside the room. This idea is the main motivation for the work presented in this thesis. In particular, this research is focused in estimating the room geometry by means of small distribute microphone arrays as this problem can be seen as fundamental towards predicting the sound field in a room. Furthermore, as one of the most common spaces for audio reproduction are shoe-box shaped rooms, the results are evaluated on this type of room. However, the methods from this thesis can, in principle, be generalized to accommodate arbitrary room shapes.

1.1 Research statement and outline

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

> *How can the boundary of a shoe-box shaped room be efficiently identified from measured data using small microphone arrays?*
>
> in two different scenarios:
>
> - Room impulse responses are known
> - Raw microphones measurements are available

The rest of the thesis continues as follows. Chapter 2 provides the problem description, highlights the importance of estimating the room geometry, and presents the contributions of this thesis. In Chapter 3 the state-of-the-art solution for the case of known RIRs is discussed and an alternative approach with lower complexity is presented as the first contribution of this thesis. Chapter 4 introduces the second contribution of this thesis as a practical solution for the problem of room geometry estimation when real-life considerations are made. Results from simulations are presented in both chapters to evaluate the proposed methods. Finally, Chapter 5 presents the conclusions and future research directions.

Problem Description 2

Estimating the sound field at an arbitrary point in an enclosure means being able to, from a set measurements, describe the emitting source and propagation environment. In typical acoustical reproduction settings, the source can be considered known. That is, the reproduced content is controlled by the user, and the specification of the drivers, i.e., loudspeakers, is known by the manufacturer. Hence, the challenge is to adequately describe the reproduction environment. Usually, the space in which the content is reproduced is considered unknown as it can greatly change from user to user.

This chapter addresses the background theory needed for the work in this thesis and a brief literature review of prior art for sound field estimation and room boundary identification. In addition, the particular research questions and contributions of this thesis are presented.

2.1 Background Theory

2.1.1 Wave Equation

Any complex sound field can, in principle, be considered as the superposition of an infinite set of simple sound waves, e.g., plane, cylindrical, spherical, etc. Furthermore, the propagation of such waves can be considered linear if the medium where they travel is homogeneous and independent of the wave amplitude [29].

The wave equation is the expression that governs the propagation of waves through fluids, i.e., gas or liquid. This equation, based on second order partial differential equations (PDE) [53], describes the evolution of the sound pressure $p(\mathbf{r}, t)$ as a function of space $\mathbf{r} = [x, y, z]^T \in \mathbb{R}^3$ and time $t \in \mathbb{R}$.

For the case of an homogenous medium with no viscosity, it is possible to linearize several relations in order to state the well-known wave equation [43]

$$\nabla^2 p(\mathbf{r}, t) - \frac{1}{c}\frac{\partial^2 p(\mathbf{r}, t)}{\partial t^2} = 0, \tag{2.1}$$

where the c is the speed of sound in the medium and ∇^2 is the Laplacian in Cartesian coordinates (x, y, z) given by

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}. \tag{2.2}$$

In real situations, inhomogeneities can occur in the medium. The most common are due to temperature changes and fluid movements due to air circulating systems. However, these perturbations are so small that they can usually be ignored.

Besides the relation of the homogenous PDE in (2.1), the source function and the boundary conditions for the PDE, describing the reflections at the walls, are needed in

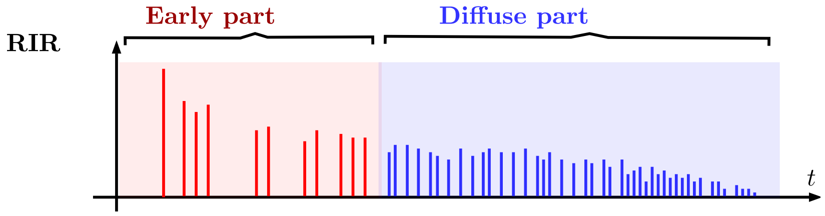


Figure 2.1: Illustration of a room impulse response

order to calculate the sound field established by a given source in a specific room. The inhomogenious wave equation, including the source function, is given by

$$\nabla^2 p(\mathbf{r},t) - \frac{1}{c}\frac{\partial^2 p(\mathbf{r},t)}{\partial t} = -s(\mathbf{r},t), \tag{2.3}$$

where $s(\mathbf{r},t)$ is the sound pressure function of the source. Notice that (2.3) cannot be solved until the boundary conditions for the PDE are completely defined.

Now, consider a source function describing an harmonic disturbance given by

$$s(\mathbf{r},t) = S(\mathbf{r};\omega)e^{j\omega t}, \tag{2.4}$$

the solution of (2.3), in the frequency domain, for arbitrary boundary conditions is given by [43]

$$P(\mathbf{r};\omega) = \int\int\int_{\mathcal{V}_s} H(\mathbf{r},\mathbf{s};\omega)S(\mathbf{s};\omega)d\mathbf{s}, \tag{2.5}$$

where $H(\mathbf{r},\mathbf{s};\omega)$ is the room transfer function (RTF) (Green's function), $\mathcal{V}_s$ denotes the source volume, $d\mathbf{s} = (dx_s, dy_s, dz_s)$ is the differential volume element, and $\mathbf{s}$ is the position of the differential contribution. The time domain counterpart of the solution $p(\mathbf{r},t)$ can be found by computing the inverse Fourier transform of (2.5).

The main challenge of sound field estimation, when the source function is known, comes from the fact that $H(\mathbf{r},\mathbf{s};\omega)$ is not only dependent on the position of the source $\mathbf{s}$, but also from the position of interest (POI) $\mathbf{r}$. As there is no reason to believe that, in an arbitrary enclosure $\mathcal{V}_R$, the set of RTF $\mathcal{H}_\omega = \{H(\mathbf{r},\mathbf{s};\omega)\}_{\mathbf{r}\in\mathcal{V}_R}$ has a particular structure, trying to estimate $H(\mathbf{r}^*,\mathbf{s};\omega)$ from a set $\tilde{\mathcal{H}}_\omega \subset \mathcal{H}_\omega$, which does not contain $H(\mathbf{r}^*,\mathbf{s};\omega)$, can prove itself a challenging task.

In order to avoid this problem, in this work attention is given to the image source model [1]. This model provides a reasonable parametrization of the RTFs based on the positions of the boundaries in the case of a room with flat walls. It considers reflections to be specular and that reflection coefficients are frequency-independent. In addition, air absorption is neglected. Hence, the spatial information from the early part of the RIR (see Fig. 2.1) is properly maintained. However, as it only considers specular reflections, at high frequencies, where objects in the room have similar size to the wavelengths of sound, and diffraction occurs, the method fails to accurately represent the sound propagation. Notice that this geometrical interpretation for acoustic propagation is only valid for the limiting case of vanishingly small wavelengths [29]. This condition is usually met in acoustics when the dimensions of the listening room and its walls are large compared with the wavelength of sound. The vanishingly small wavelengths assumption often holds, in typical rooms, at frequencies larger than 1000Hz [29].

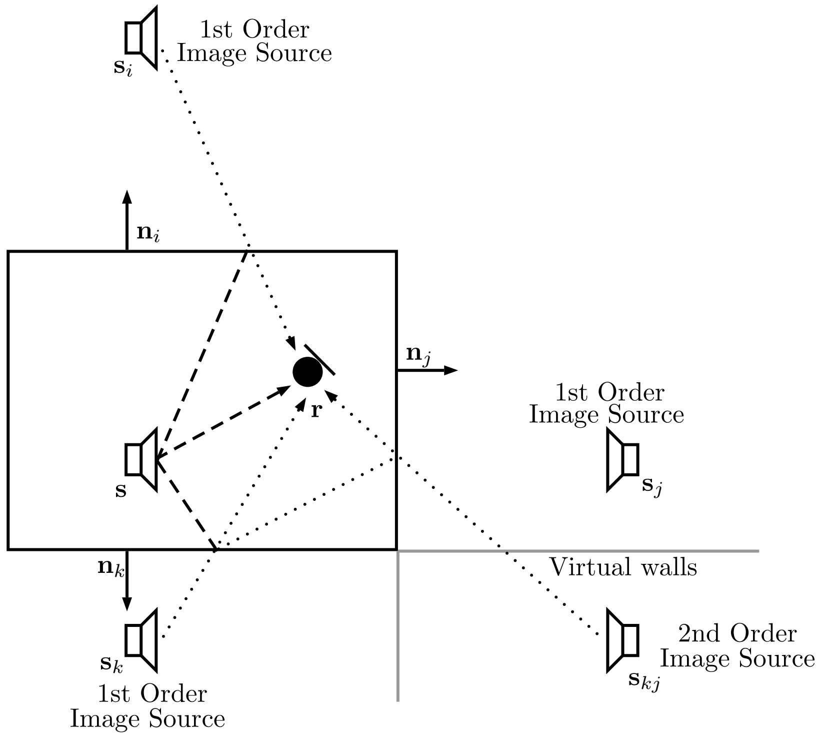


Figure 2.2: Illustration of the image source model for a room. By mirroring the original source with respect the walls, the image sources are defined

In the following section, the image source model and the induced RTFs parametrization are properly introduced.

2.1.2 Image Source Model

The main idea behind the model proposed by Allen and Berkley in [1] is that the reflections in a room can be interpreted as contributions of virtual sources located at positions that provide an equivalent propagation's path length. That is, the image source model allows us to replace a reflection coming from a wall by a virtual source mirrored in the boundaries of the room. Fig. 2.2 illustrates the creation of equivalent image sources by mirroring the original source with respect to the walls. As shown in the same figure, higher order reflections are result of the mirroring of other image sources in the generated virtual walls.

Consider the time delay of arrival (TOA) at a location $\mathbf{r}$ of a sound ray from the source s be given by

$$\tau = \frac{d}{c} = \frac{\|\mathbf{s} - \mathbf{r}\|_2}{c}, \tag{2.6}$$

where $\mathbf{s} \in \mathbb{R}^3$ is the position of the source s.

For an omnidirectional point source s in free space, with Green's function given by [53]

$$H_{free}(\mathbf{r}, \mathbf{s}; \omega) = \frac{\exp(j\omega\tau)}{4\pi\|\mathbf{r} - \mathbf{s}\|_2}, \tag{2.7}$$

the image source model provides a RTF given by

$$H(\mathbf{r},\mathbf{s};\omega) = H_{free}(\mathbf{r},\mathbf{s};\omega) + \sum_{i\in\mathcal{I}} \gamma_i \frac{\exp(j\omega\tau_i)}{4\pi\|\mathbf{r}-\mathbf{s}_i\|_2}, \tag{2.8}$$

where $\mathcal{I}$ is the set of reflections considered in the model, s_i the i-th reflection of source s, τ_i is the associated TOA of the reflection s_i, and γ_i is the attenuation coefficient related with the i-th wall.

Considering $s_0 \triangleq s$, $\gamma_0 = 1$, and taking the inverse Fourier transform of (2.8), the time domain counterpart of the RTF, the RIR, is obtained

$$h(\mathbf{r},\mathbf{s},t) = \sum_{i=0}^{|\mathcal{I}|} \gamma_i \frac{\delta(t-\tau_i)}{4\pi\|\mathbf{r}-\mathbf{s}_i\|_2}. \tag{2.9}$$

The RIR in (2.9) can be seen as a train of pulses, each corresponding to a delayed version of the original source s, i.e., reflection. Notice that in this model only the attenuation coefficient, associated with the surface material of the walls, and the distance between source position and POI affect the amplitude of the delta function.

The resulting filter from (2.9) modifies any sound emitted in the room. At any arbitrary location $\mathbf{r}$, the observed sound pressure $p(\mathbf{r},t)$ is the result of the convolution of the original source function $s(t)$, located at position $\mathbf{s}$, with the RIR, i.e.,

$$p(\mathbf{r},t) = s(t) * h(\mathbf{r},\mathbf{s},t). \tag{2.10}$$

Even though the model in (2.10) still depends on the POI $\mathbf{r}$, the image source model provides a parametrization of the RTF (RIR) based on image sources, which are directly obtained from the room walls. For example, a first-order reflection coming from the j-th wall can be expressed in terms of the source location and the wall it is mirrored in by

$$\mathbf{s}_j = \mathbf{s} + 2\langle \mathbf{p}_j - \mathbf{s}, \mathbf{n}_j\rangle \mathbf{n}_j, \tag{2.11}$$

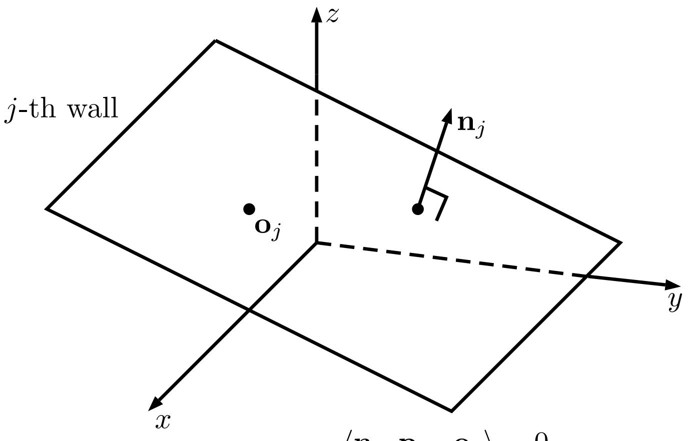
where $\mathbf{s}_j$ is the position of the first-order image source corresponding to the j-th wall, $\mathbf{p}_j$ is an arbitrary point on the j-th wall, and $\mathbf{n}_j$ is normal pointing outward from the room with respect the j-th wall. Hence, the RIR in (2.9) can be described for any arbitrary position $\mathbf{r}$ given that the position of the source and the walls are known. This is the reason why, as first step towards sound field estimation, this work focuses on the estimation on the boundaries of the room using the image source model.

2.1.3 Room Reconstruction

An important property of the image source model is its geometrical duality. In the same way that the reflections can be found from the source and walls locations, the walls can be found from the positions of the image sources.

Let $\mathbf{s}$ denote the location of the source and $\mathbf{s}_j$ the location of the first-order image source with respect to the j-th wall. The normal vector of the j-th wall is given by

$$\mathbf{n}_j = \frac{\mathbf{s}-\mathbf{s}_j}{\|\mathbf{s}-\mathbf{s}_j\|_2}. \tag{2.12}$$



z
j-th wall
$\mathbf{n}_j$
$\mathbf{o}_j$
y
x
$\langle \mathbf{n}_j, \mathbf{p} - \mathbf{o}_j \rangle = 0$

Figure 2.3: Boundary plane of the j-th wall

By using the normal vector $\mathbf{n}_j$ and a point from the j-th wall given by

$$\mathbf{o}_j = (\mathbf{s} + \mathbf{s}_j)/2, \tag{2.13}$$

the boundary plane from the j-th wall can be reconstructed. The equation of this plane given by

$$\langle \mathbf{n}_j, \mathbf{p} - \mathbf{o}_j, \rangle = 0, \tag{2.14}$$

describes the points $\mathbf{p}$ found on the wall plane shown in Fig. 2.3.

The vertices of the room can be found as the intersections of the boundary planes. Generally, the accuracy in the estimation of the position of the j-th wall increases as more points of the boundary plane are estimated. These points can be generated by either using more sources or moving the original source inside the enclosure.

2.2 Prior Art

2.2.1 Sound Field Estimation

The idea of sound field estimation approaches the fundamental problem of transfer function interpolation/extrapolation [21]. For example, consider the situation shown in Fig. 2.4. A set of loudspeakers (gray circles) measure the sound emitted by a TV (black rectangle) using the microphones mounted on them. In this situation, it is desirable to know how the sound propagates from the TV to the listener position (region of interest) to enhance the listening experience. Therefore, the sound field at the listener location has to be estimated using the measurements of the sound field acquired at the positions of the microphones. This problem can be posed as one that uses the RIRs at the microphones locations, estimated through measurements, to create an estimate of the RIR that describes the sound propagation from the source to the listener location. This problem is of high interest in the field of acoustics as this information could improve user experience [4], i.e., if the transfer function from the source is known for every point in a room, corrections for the room acoustics can be made in order to enhance reproduction quality at any position.

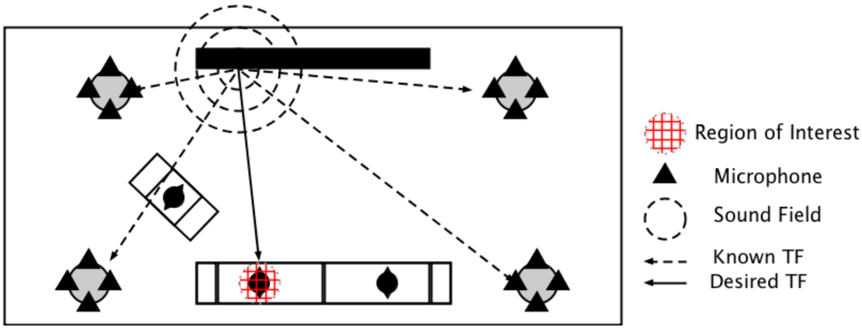


Figure 2.4: Illustration of the problem of estimating the transfer function (TF) of a region of interest given a set of other TFs outside it

Classic approaches physically sample the space in dense clusters to compute such transfer functions in different spatial regions, leading to usage of expensive equipment and/or a time consuming process [38]. These methods are highly dependent on how dense the sampling is, rendering the approach useless for a sparse distributed network of sensors as the one shown in Fig. 2.4. Different methods, spanning from classic filter theory, have been proposed in the past to deal with sparse measurements, however it has been shown that acceptable results can only be achieved when the point of interest is found near the measurements positions [37]. Hence, this kind of approach becomes unfit for large spaces as the ones found in typical audio reproduction setups. In addition, those methods assume a reference signal at the desired position to fit a model, which is equivalent to perform a survey of the sound field at the point of interest. Recently, the estimation problem has been approached through arbitrary array geometries by means of an equivalent source model [8][15][36]. Under this model, common for source imaging [4], the problem is being solved by claiming a perceptual equivalence with respect to the original sound signal. Such methods limit the scope of the application as their outputs are perceptual equivalent signals, and the sound field is not properly estimated. That is, the output of these kind of methods does not represent the actual physical state of the field, i.e., physically correct pressure values at the points of interest. These results remove the possibility of driving a secondary source to interact, by constructive or destructive interference, with the field present at the desired position.

Besides classical methods, recent methods have tried to combine the strength of data-driven and model-based approaches [50][51][3]. These methods lever on the measurements acquired at the receivers locations and constraint the reconstructed field with a physical model described in terms of a partial differential equation (PDE). These approaches result in the minimization of a cost function under linear constraints. Even though these kind of methods promise good results, they suffer from the same sampling problems as the classical approaches. As they are based on finite elements or finite differences, they require a dense grid of measurements in order to provide good estimates for complex fields.

2.2.2 Room Boundary Estimation

Considering the limitations of the methods for estimating directly the sound field, this thesis addresses the problem of room geometry. Instead of directly trying to predict the sound pressure at a given position, the parametric description of the propagation of sound, given by the image source model, is employed for finding the room geometry. By finding the locations from which the reflections occur in a room, it is possible to simulate how the sound propagates in the enclosure until it arrives at the point interest. Hence, the work in this thesis is mainly concerned with estimating the position of such boundaries.

In the literature, there are several methods available for obtaining the shape of an enclosure. Most of those methods assume knowledge of the room impulse responses. In [10], the shape in the 2D case is estimated by a single RIR. Antonacci et al. [2] solve the 2D problem assuming multiple sources and microphones. Dokmanić et al. in [12] exploits the properties of Euclidean distance matrices (EDMs) to find the room geometry in the general 3D case. More recently, a newly proposed method [25] by Jager et al. has been shown to provide the same accuracy as Dokmanić's method at a much lower computational complexity. This approach recasts the labeling problem of the acoustic echoes problem into a graph problem.

Even though the state-of-the-art methods for room geometry estimation provide an accurate solution to the problem, they suffer from two main issues: (*i*) high computational complexity and (*ii*) dependency of an *oracle* capable of identify the peaks of the RIRs that represent proper reflections. Motivated by these issues, this thesis aims to (*i*) reduce the computational complexity of the state-of-the-art methods by exploiting the subspace properties of the data model, and (*ii*) to devise methods capable of being used in real applications where, most of the time, RIRs are not available and general methods capable to use arbitrary signals are desired. Even though the room shape estimation problem can be addressed for arbitrary room geometries, the work is restricted to shoe-box shaped rooms as they are commonly found in typical audio reproduction situations. In addition, as we aimed to develop methods applicable in audio reproduction systems, consisting of multiple loudspeakers with built-in microphones, our main interest is to explore up to which extend small microphone arrays can be employed to estimate such boundaries. This auto-imposed constraint sets the general research question as a feasibility study rather than a designing task. That is, differently from solutions that could be devised to solve the same problem for controlled situations, i.e., acoustic consulting with ad hoc instruments, this project is focused on the ability of a small set of microphone arrays to solve the room geometry estimation problem.

2.3 How to efficiently find the first-order reflections?

So far the challenging task of sound field estimation has been discussed, and the importance of the identification of the room boundary for this problem highlighted. However, the question of how to find the first-order reflections needed for room reconstruction, using measured data, has not been addressed.

This work aims to develop efficient methods for identifying first-order reflections capable of delivering equivalent performance compared to current state-of-the-art methods while reducing the computational complexity involved. Furthermore, as in actual sound systems the estimation of the first-order reflections becomes even more challenging, the part of our interest resides on robust methods capable to deal with complications that might appear in real situations, i.e., loudspeaker transfer functions, absence of oracle, impossibility of measuring RIRs, etc.

As a result, the particular questions addressed in this thesis include:

- (*i*) Is it possible to find a faster alternative to the graph-based strategy for acoustic echoes sorting problem? If so, can we guarantee the same estimation performance?
- (*ii*) When the RIRs are not available, or the TOA estimates for all microphone-image source pair cannot be properly identified, is it possible to devise an estimator capable to find the first-order reflections from the raw microphone measurements?

Therefore, in this thesis, the following contributions towards solving the problem of estimating the room geometry for sound field estimation using a set of distributed microphone arrays are made:

- A fast greedy subspace-based algorithm for efficient echo labeling and source localization in the case of known RIRs
- Sequential iterative algorithms based on a non-linear cost function capable of estimating the room geometry in the case of raw measurements, possibly from arbitrary transmitted signals

In theory, going from the raw measurements to known RIR should be possible. However, there is no guarantee that the signal used during transmission is able to provide a proper estimation of the RIRs. Furthermore, the assumptions made in the acoustic echoes sorting problem do not necessary hold in all circumstances. Both situations, though closely intertwined, are in principle different and require distinct approaches in order to extract the first-order reflections positions form the available data. As a result, this thesis focuses on the *how to* for finding the reflections in each scenario.

Acoustic Echo Sorting for Source Localization 3

The aim of this chapter is twofold: (*i*) to introduce the current state-of-the-art solution for the acoustic echo sorting problem for shoe-box shaped rooms, and (*ii*) to present the first contribution of this thesis. The contribution is a subspace-based method for acoustic echoes sorting which provides the same accuracy as the state-of-the-art solution at a reduced computational complexity. This is achieved by omitting an NP-hard graph problem through a greedy strategy, and using a subspace condition to reduce the number of feasible combinations of echoes. In addition, when the positions of the sources, i.e., first-order reflections, are estimated using known microphones locations, the proposed method only requires measurements from a single source. This contrasts with the current state-of-the-art approach which requires more than one source in order to disambiguate its results.

3.1 Acoustic echo labeling problem

As explained in the previous chapter, in order to estimate the boundaries of a shoe-box shaped room, the source and the six first-order reflections, i.e., four walls, floor, and ceiling, must be located. Even though the estimation of the source locations from known labeled TOAs could be seen as a straightforward problem, the labeling of the TOAs, i.e., relating the peaks in different RIRs with a unique boundary, is not. This challenge arises from the fact that reflections can arrive in different order at the microphones locations. This ambiguity issue is illustrated in Fig. 3.1.

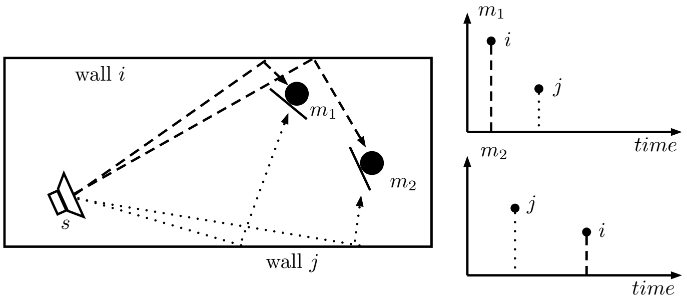


Figure 3.1: Example for the, possible, different order of arrival of boundary reflections

In order to solve this problem, this chapter deals with an instance of the acoustic echo labeling problem where the following assumptions hold:

- (*A.1*) **Oracle**
  From given RIRs, it is possible to identify the peaks in the response corresponding to the room boundaries, i.e., TOAs from the reflections are always available.

- (*A.2*) **Synchronization**
  It is assumed that either the TOAs are absolute, i.e., the microphones and sources are precisely synchronized, or that it is possible to obtain absolute TOAs by means of least squares.

- (*A.3*) **TOAs Accuracy**
  It is possible to estimate the TOAs up to an arbitrary accuracy σ^2_{TOA}.

- (*A.4*) **Known Source Position**
  The source position is either known a priori, or it is assumed that is possible to localize it by trilateration using the first peak of the different RIRs.

- (*A.5*) **Known Microphones Positions**
  The relative position between microphones is known up to a rigid transformation, i.e., translation, rotation, etc.

From these set of assumptions, the most restrictive in practice is (*A.1*). However, for well-behaved rooms it is possible to assume that the peaks can be easily found by selecting the highest peaks in the RIRs. For cases in which (*A.1*) is not a feasible assumption, or the peak picking problem proves itself harder than the acoustic echo labeling problem, the next chapter provides an alternative approach to room geometry estimation. For the rest of the assumptions, there are either methods already proved in the literature to deal with the problems or they can be guaranteed by system design. For example, in the case of (*A.2*) either a single interface is used for microphones and loudspeakers (perfect synchronization) or the offset in the measured RIRs can be estimated. Assumption (*A.3*) can be considered to hold in most of the instances as the uncertainty in the estimation method used for finding the TOAs can be known a priori [45]. As it is considered that there are no peaks before the one corresponding to the direct path, the assumption (*A.4*) always holds. The relative positions of the microphones (*A.5*) can either be known by construction of the measuring setup or using state-of-the-art methods available in the the literature [40][42][7].

Considering this, the echo sorting problem, after converting TOAs into distances, can be defined as follow:

> **Echo Sorting Problem** (*P.1*)
> From a set $\mathcal{D}$ containing the unlabeled squared distances between the M microphones and the N sources, obtain the distance matrix $\mathbf{D} \in \mathbb{R}^{M \times N}$, where the n-th column contains the squared distances between the M microphones and the n-th source $\forall\, n \in 1, \ldots, N$.

To illustrate how this problem can be approached, let us consider the following set of unlabeled squared distances

$$\mathcal{D} = \{d_{mn}\} \;\forall\; (m, n) \in [1, \ldots, M] \times [1, \ldots, N], \tag{3.1}$$

where $|\mathcal{D}| = N^M$ and d_{mn} represents the squared distance between the m-th microphone and the n-th source. As the subindex n is *hidden*, i.e., the source or reflection

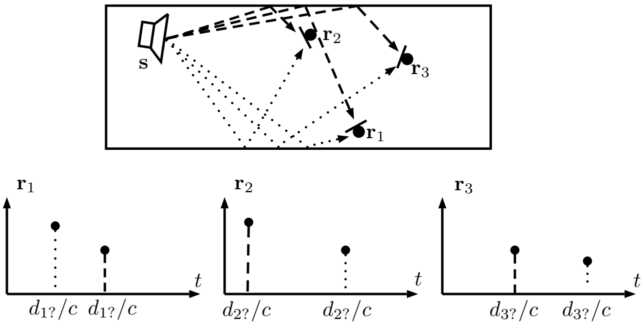


Figure 3.2: Ambiguity in the sorting of the received echos

which originated the echo is unknown, all the possible echoes combinations need to be generated to find the correct combination. This process leads to a set of N^M possible combinations. Notice that the definition employs the squared of the distances instead of the Euclidean distance. This choice becomes clear in the following sections when the EDMs are introduced (see Appendix A).

For the sake of clarity, consider the case of $M = 3$ and $N = 2$ (see Fig. 3.2). All the possible combinations for this instance listed in matrix form, where each column represents an echo combination, are given by

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

In this situation, a set of TOAs is available, but it is ignored to which boundary they belong. Hence, a strategy to group the echoes accordingly to the wall that originated them is required. Following this idea, the problem (*P.1*) can then be solved by finding a strategy that retrieves the N columns of $\tilde{\mathbf{D}}$ which contains the true echo combinations. In this particular example, the correct combinations are the columns $\{1, 8\}$. It is important to point out that the matrix in (3.2) is just an instance of the combination generation process. The reader must not be confused with the fact that in this example the true combinations are in the first and last columns. In a real problem instance, when $\tilde{\mathbf{D}}$ is generated from $\mathcal{D}$, $\tilde{\mathbf{D}}$ becomes a shuffled version, in the columns, of (3.2).

In the following sections, the state-of-the-art method to solve the problem and the first contribution of the thesis are introduced as column picking strategies.

3.2 Graph-based (state-of-the-art) approach

The current state-of-the-art method for column picking from $\tilde{\mathbf{D}}$ was recently proposed by Jager et al. in [25]. This approach is a fast alternative to the method proposed by Dokmanić et al. in [12] based on multidimensional scaling (MDS). It provides the same estimation accuracy as Dokmanić method at a reduced computational cost. In this approach the rank constraint of the EDMs is exploited to reduce the number of combinations that should be considered. Furthermore, when the set of combinations has been reduced, a key observation is made: it is very unlikely that two feasible combinations share elements in common. This observation is used to shown that it is possible to recast the original problem into a graph problem where the maximum independent sets of a graph lead to a tractable strategy to select the columns from $\tilde{\mathbf{D}}$.

As it is not intended in this work to give an in-depth treatment of Jager's method, the overall flow of this approach for echo labeling is summarized in the following steps:

- **Pre-filtering**
  Following the result of Theorem 1 in Appendix A, an EDM $\mathbf{E}$ with affdim($\mathbf{E}$) = 3 (see Definition 1 in Appendix A), has at most a rank of 5. Consider $\mathbf{E} \in \mathbb{EDM}$, where $\mathbb{EDM}$ represents the set of all EDMs, be built by using the squares of the relative distances between microphones, i.e.,

  $$\mathbf{E} = \begin{bmatrix} d_{\mathbf{r}_1\mathbf{r}_1} & \cdots & d_{\mathbf{r}_1\mathbf{r}_M} \\ \vdots & \ddots & \vdots \\ d_{\mathbf{r}_M\mathbf{r}_1} & \cdots & d_{\mathbf{r}_M\mathbf{r}_M} \end{bmatrix} \in \mathbb{R}^{M\times M}, \tag{3.3}$$

  where $d_{\mathbf{r}_i\mathbf{r}_j}$ denotes the squared Euclidean distance between the i-th and j-th microphones. Using Theorem 1, the feasibility (as a true echo combination) of the c-th column of $\tilde{\mathbf{D}}$ can be tested by forming the augmented matrix

  $$\mathbf{E}_c = \begin{bmatrix} \mathbf{E} & \tilde{\mathbf{D}}_c \\ \tilde{\mathbf{D}}_c^T & 0 \end{bmatrix} \in \mathbb{R}^{(M+1)\times(M+1)}, \tag{3.4}$$

  where $\tilde{\mathbf{D}}_c \in \mathbb{R}^{M\times 1}$ denotes the c-th column of $\tilde{\mathbf{D}}$, and checking the rank constraint for EDMs over $\mathbf{E}_c$. In practice, due to the uncertainties in the TOA estimates, it is not realistic to apply a hard threshold over the columns of $\tilde{\mathbf{D}}$. Instead, in [25] the ϵ-rank [16], which is defined by

  $$\operatorname{rank}(\mathbf{E}_c, \epsilon) = \min_{\|\mathbf{E}_c - \mathbf{X}\|_2 \le \epsilon} \operatorname{rank}(\mathbf{X}), \tag{3.5}$$

  is used as a realistic surrogate. This condition filters out singular values from the SVD lower than ϵ, which most probably are due to perturbations in the measurements. In the case of noiseless measurements this step provides a perfect column picking strategy. However, as this is not realistic in practice, the pre-filtering step generates a set of indexes for the columns of $\tilde{\mathbf{D}}$ given by

  $$\mathcal{C}_\epsilon = \{c : \operatorname{rank}(\mathbf{E}_c, \epsilon) \le 5\}, \tag{3.6}$$

  which most probably contains false positives. The output of this step can be considered as

  $$\tilde{\mathbf{D}}_{\mathcal{C}_\epsilon} \in \mathbb{R}^{M\times|\mathcal{C}_\epsilon|}, \tag{3.7}$$

  which denotes the matrix of possible echo combinations built by selecting the columns of $\tilde{\mathbf{D}}$ listed in $\mathcal{C}_\epsilon$. Note that in most the cases, when there is noise in the measurements, $|\mathcal{C}_\epsilon| \gg N$.

- **Maximum Independent Sets**
  For the cases where $|\mathcal{C}_\epsilon| > N$ a way to identify the elements that correspond to

the feasible echos combinations is required. In [25] it is noticed that the vectors containing feasible combinations are very unlikely to have elements in common. Hence, the number of columns of $\tilde{\mathbf{D}}_{\mathcal{C}_\epsilon}$ can be further reduced by selecting a subset with vectors with no elements in common. For this purpose, a simple graph $G(V, E)$ is constructed considering each column of $\tilde{\mathbf{D}}_{\mathcal{C}_\epsilon}$ as a node. Two vectors (nodes) are considered adjacent in the graph if they share at least one element in common. Extracting the maximum independent sets $\mathcal{S}^G_{max}$ (see Appendix B) from $G(V, E)$ is shown to be tantamount to selecting the columns of $\tilde{\mathbf{D}}_{\mathcal{C}_\epsilon}$ with no elements in common. As in most of the instances the cardinality of $\mathcal{S}^G_{max}$ is greater than one, an additional step is required in order to decide which set corresponds to the correct set of echo combinations.

- **Pollefey's method**
  At this point, the method requires to choose one of the sets in $\mathcal{S}^G_{max}$ to provide the correct labels for the echos. This is achieved using Pollefey's method [41] and selecting the set that provides the best fit in the reconstruction, i.e., provided the distances, Pollefey's method is able to estimate (up to a non-singular transform) the location of the sources and microphones. The fit is considered as the error between the true and estimated microphone locations after a Procrustes analysis [17]. For this step, due to the computationally cost of the previous step, i.e. the number of microphones are restricted, in [25] the usage of $Q \geq 2$ sources is required in order to meet the requirements of Polleyfey's method. This step requires to try all combinations of the sets in $\{\mathcal{S}^{G_1}_{max}, \ldots, \mathcal{S}^{G_Q}_{max}\}$.

At the end of the processing chain, the correct echoes labels and sources (first-order reflections) locations are obtained. With this information is then possible to reconstruct the room boundary by straightforward geometrical methods as described in Chapter 2. The graph-based approach proposed by Jager et al. provides a fast alternative for solving the echo sorting problem compared with the method of Dokmanić et al. using MDS. While the method based on MDS could take hours to run for small instances of M and N, the graph-based method only requires several seconds. This is a large improvement in terms of computational cost without compromising accuracy in the estimation. However, as the maximum independent set problem is a NP-hard problem, in instances where the pre-filtering stage, based in the ϵ-rank constraint, is not able to considerably reduce the number of feasible combinations the graph problem becomes intractable. These large instances of the problem can arise when the distances of echoes are not properly estimated, i.e., high uncertainty. Hence, a different approach capable to deal with high uncertainties or a strategy able to further reduce the feasible combinations would of interest. In the next section, an alternative method, based on a greedy strategy, is introduced in order to alleviate the problem's complexity.

3.3 Subspace-based (Greedy) Approach

In this section, an alternative solution to the acoustic echo labeling problem is presented. Similarly to the graph-based approach, a filtering strategy capable to perfectly select the correct columns of $\tilde{\mathbf{D}}$ under noise-free conditions is proposed. For the case of uncertainty in the TOAs estimates, a greedy strategy is employed to avoid the maximum independent set problem of the graph-based approach.

Consider an arbitrary set $\mathcal{M}$ of M receivers located at random positions. That is, $\mathcal{M} = \{\mathbf{r}_m = [x_m, y_m, z_m]^T \in \mathbb{R}^3\}_{m=1}^{M}$. These locations are considered known up to a non-singular transformation. Furthermore, consider the set $\mathcal{S} = \{\mathbf{s}_n = [X_n, Y_n, Z_n]^T \in \mathbb{R}^3\}_{n=1}^{N}$ of N sources. The squared distances $\mathcal{D} = \{d_{m,n}\} \forall\, (m,n) \in [1,\dots,M] \times [1,\dots,N]$ between the sources $\mathcal{S}$ and receivers $\mathcal{M}$ are assumed to be known, i.e., the time-of-arrival can be estimated at the receivers. Hence, the squared distance $d_{m,n}$ for the (m,n)-th pair can be written as

$$(x_m - X_n)^2 + (y_m - Y_n)^2 + (z_m - Z_n)^2 = d_{m,n}. \tag{3.8}$$

This can be expressed in a vector notation as [41]

$$\mathbf{R}_m^T \mathbf{S}_n = d_{m,n}, \tag{3.9}$$

where

$$\mathbf{R}_m = [\mathbf{r}_m^T \mathbf{r}_m \ -2x_m \ -2y_m \ -2z_m \ 1]^T \in \mathbb{R}^{5\times 1}, \tag{3.10}$$

$$\mathbf{S}_n = [1 \ X_n \ Y_n \ Z_n \ \mathbf{s}_n^T \mathbf{s}_n]^T \in \mathbb{R}^{5\times 1}. \tag{3.11}$$

Collecting all the squared distances $d_{m,n}$ for the pairs (m,n) leads to the distance matrix $\mathbf{D} \in \mathbb{R}^{M\times N}$, and the model can be written in matrix form as

$$\mathbf{R}^T \mathbf{S} = \mathbf{D} \in \mathbb{R}^{M\times N}, \tag{3.12}$$

where $\mathbf{R} = [\mathbf{R}_1 \dots, \mathbf{R}_M]$ and $\mathbf{S} = [\mathbf{S}_1, \dots, \mathbf{S}_N]$ are the microphone and source positions matrices respectively. Even when the positions of the microphones are known up to an arbitrary non-singular matrix $\mathbf{H} \in \mathbb{R}^{5\times 5}$, and $\hat{\mathbf{R}}^T = \mathbf{R}^T\mathbf{H}$ is known instead of $\mathbf{R}$, the model in (3.12) still holds as

$$\hat{\mathbf{R}}^T \mathbf{H}^{-1} \mathbf{H} \hat{\mathbf{S}} = \mathbf{D}, \tag{3.13}$$

where $\hat{\mathbf{S}} = \mathbf{H}^{-1}\mathbf{S}$ is the transformed matrix of sources positions.

From the model in (3.12), when the positions of the receivers and the distance matrix $\mathbf{D}$ are known, the only unknown is the matrix $\mathbf{S}$ with the position of the sources. This problem could be solved by means of least squares given that $M \geq 5$. However, as instead of (3.12) the available matrix is $\tilde{\mathbf{D}}$, which contains all possible combinations of the distances in $\mathcal{D}$, the modified data model is given by

$$\begin{aligned} \tilde{\mathbf{D}} &= [\mathbf{R}^T \mathbf{S} \ \mathbf{A}] \mathbf{P}_\Pi & (3.14) \\ &= [\mathbf{D} \ \ \mathbf{A}] \mathbf{P}_\Pi \in \mathbb{R}^{M\times N^M}, & (3.15) \end{aligned}$$

where $\mathbf{A} \in \mathbb{R}^{5\times(N^M - N)}$ is an arbitrary matrix with unknown structure and $\mathbf{P}_\Pi^T \in \mathbb{R}^{N^M \times N^M}$ is a random permutation matrix which shuffles the columns of $\tilde{\mathbf{D}}$ allocating in its first N columns the true echo combinations.

From the model in (3.15) several strategies to identify $\mathbf{D}$ can be devised. For example, a straightforward approach, considering $M \geq 5$, could use the pseudoinverse of $\mathbf{R}^T$ to find the correct combinations, i.e.,

$$(\mathbf{R}^T)^\dagger \tilde{\mathbf{D}} = (\mathbf{R}^T)^\dagger [\mathbf{D} \;\; \mathbf{A}]\mathbf{P}_\Pi \;\; = \;\; [\mathbf{S} \;\; (\mathbf{R}^T)^\dagger \mathbf{A}]\mathbf{P}_\Pi, \tag{3.16}$$

where $(\mathbf{B})^\dagger$ is the Moore-Penrose pseudoinverse of $\mathbf{B}$. Using the structure of the matrix $\mathbf{S}$, defined by (3.11), it is possible to discard columns from $\tilde{\mathbf{D}}$ which when multiplied by $(\mathbf{R}^T)^\dagger$ do not meet the following constraints:

- (i) First element of the column equal to one
- (ii) Last element of the column has to be positive

However, in most of the cases the pseudoinverse of $\mathbf{R}^T$ behaves as an expansive operator, i.e., $\|(\mathbf{R}^T)^\dagger\|_2 > 1$, potentially increasing any existing measurement noise. In addition, due to the unknown structure of the matrix $\mathbf{A}$, the conditions (i) and (ii) are only *necessary conditions* (in the noise free case) for identifying columns of $\mathbf{D}$.

Therefore, a method that exploits the structure of the space spanned by the columns of $\mathbf{R}^T$ is proposed to estimate $\mathbf{D}$ from the unsorted data $\mathcal{D}$. Assuming proper diversity in $\mathbb{R}^3$ for the receivers positions, i.e., non co-located receiver positions, the only constraint needed in the method to ensure the rank-5 property of the distance matrix $\mathbf{D}$ is $M \geq 5$ [41][18].

3.3.1 Subspace Filtering

Let the SVD of the receivers position matrix $\mathbf{R}$ be given by

$$\mathbf{R} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T. \tag{3.17}$$

With this, the complementary orthogonal projection $\Pi_{\mathbf{R}}^\perp$ into $ker(\mathbf{R})$ can be computed from the SVD in (3.17) as

$$\Pi_{\mathbf{R}}^\perp = \mathbf{I}_M - \tilde{\mathbf{V}}\tilde{\mathbf{V}}^T, \tag{3.18}$$

where $\tilde{\mathbf{V}} \in \mathbb{R}^{M \times 5}$ is the economy-size $\mathbf{V}$ matrix from the SVD of $\mathbf{R}$. This projection can be shown to hold the following property

$$\Pi_{\mathbf{R}}^\perp \mathbf{D} = \Pi_{\mathbf{R}}^\perp \mathbf{R}^T \mathbf{S} = \mathbf{0}, \tag{3.19}$$

which can be used to estimate $\mathbf{D}$ from $\mathcal{D}$.
An interesting property of the complementary projection matrix is that

$$\|\Pi_{\mathbf{R}}^\perp\|_2 = 1, \tag{3.20}$$

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

$$\begin{aligned}
\|\Pi_{\mathbf{R}}^\perp(\mathbf{D} + \mathbf{N})\|_2 &= \|\Pi_{\mathbf{R}}^\perp(\mathbf{R}^T\mathbf{S} + \mathbf{N})\|_2 && (3.21)\\
&= \|\Pi_{\mathbf{R}}^\perp \mathbf{N}\|_2 && (3.22)\\
&\leq \|\mathbf{N}\|_2 && (3.23)\\
&= \sigma_{\max}(\mathbf{N}), && (3.24)
\end{aligned}$$

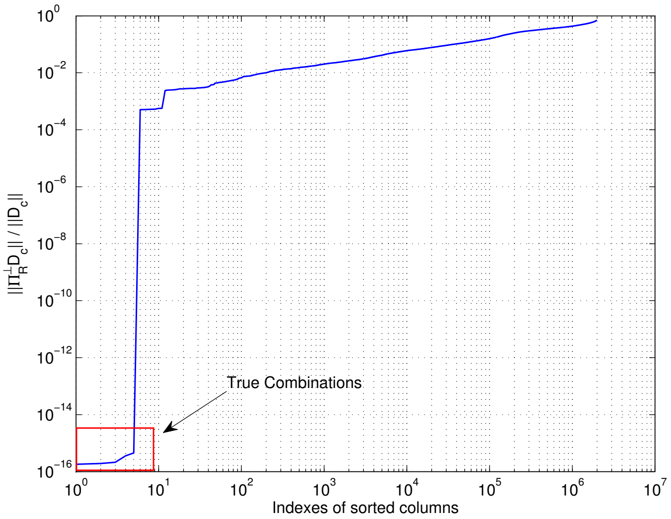


Figure 3.3: Normalized functional (3.26) for an instance of the columns $\tilde{\mathbf{D}}$ in the noise free case sorted in ascending order. In this example $M = 9$ and $N = 5$

where $\sigma_{\max}(\mathbf{N})$ is the maximum singular value of noise matrix $\mathbf{N}$. This makes the projection particularly useful in cases where the elements of $\mathcal{D}$ are perturbed with noise.

In order to apply the projection given in (3.18), consider the matrix $\tilde{\mathbf{D}}$, described in (3.2), as the distance matrix generated by all the possible combinations of the elements in $\mathcal{D}$, e.g.,

$$\tilde{\mathbf{D}} = \begin{bmatrix} d_{1,1} & \cdots & d_{1,2} & \cdots & d_{1,N} \\ d_{2,1} & \cdots & d_{2,1} & \cdots & d_{2,N} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ d_{M,1} & \cdots & d_{M,2} & \cdots & d_{M,N} \end{bmatrix} \in \mathbb{R}^{M \times N^M}. \tag{3.25}$$

In the ideal case, i.e., perfect measurements free from noise, the results are straightforward. By defining the functional

$$f(c) = \|\Pi_{\mathbf{R}}^{\perp} \tilde{\mathbf{D}}_c\|_2^2 \ \forall \ c \ \in \ [1, \ldots, N^M], \tag{3.26}$$

with $\tilde{\mathbf{D}}_c$ denoting the c-th column of $\tilde{\mathbf{D}}$, the subset of feasible columns is given by

$$\mathcal{C} = \{c \, : \, f(c) = 0\}. \tag{3.27}$$

It provides an estimate of the feasible distance matrix given by

$$\hat{\mathbf{D}} = \tilde{\mathbf{D}}_{\mathcal{C}} \in \mathbb{R}^{M \times N}, \tag{3.28}$$

where $\tilde{\mathbf{D}}_{\mathcal{C}}$ represents the trimmed distance matrix which only retains the columns specified by the set $\mathcal{C}$. The functional is illustrated in Fig. 3.3 for a problem instance with $M = 9$ microphones and $N = 5$ sources. From this figure the region where the true combinations are found can be clearly seen. In the noise-free case the knee of the graph is perfectly identified.

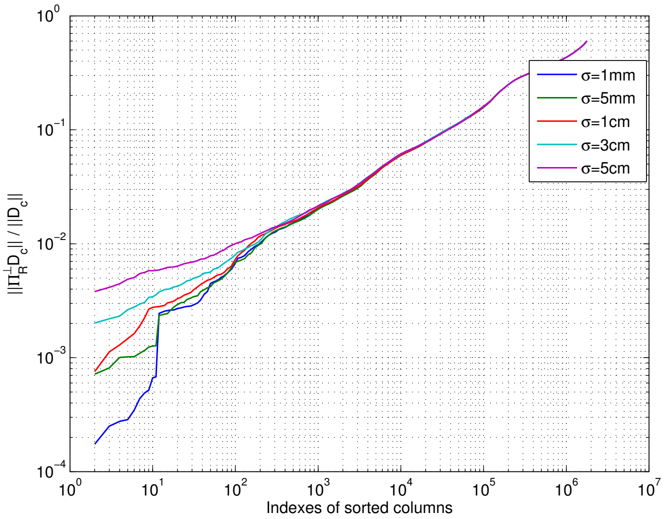


Figure 3.4: Normalized functional (3.26) for different noise levels in distances between sources and microphones

However, in real applications there is no guarantee that the true distances $\mathcal{D}$ are measured perfectly, hence the set in (3.27) is, most likely, the empty set. In order to deal with noisy measurements, a column-dependent upper bound for the proposed functional, which considers the effect of perturbations, is provided. The effect of noise in the measured distance for the proposed functional is illustrated in Fig. 3.4. Differently from the noise-free case, under the presence of noise in the measurements the knee in the plots of the functional becomes less pronounced. This problem poses natural limitations to the application of the functional.

Let us illustrate this. Consider the complementary projection being applied to the c-th column of measured data $\hat{\mathbf{D}}$ containing additive noise. The norm of this functional is bounded similarly as in (3.24). That is,

$$\|\Pi^{\perp}_{\mathbf{R}^T}\hat{\mathbf{D}}_c\|_2 = \|\Pi^{\perp}_{\mathbf{R}^T}(\mathbf{D}_c + \mathbf{N}_c)\|_2 \leq \|\mathbf{N}_c\|_2. \tag{3.29}$$

This column dependent bound becomes clear when we realize that, originally, what is estimated is the TOAs and not the squared distances $\mathcal{D}$. Consider a measurement of the TOA $\hat{\tau}_{m,n}$, with uncertainty σ_{TOA} $(A.3)$, i.e.,

$$\hat{\tau}_{m,n} = \tau_{m,n} \pm \sigma_{TOA}, \tag{3.30}$$

where $\tau_{m,n}$ is the true TOA. Transforming this quantity into the squared distance as

$$\begin{aligned} \hat{d}_{m,n} &= (c\hat{\tau}_{m,n})^2 = (c\tau_{m,n})^2 \pm 2c^2\tau_{m,n}\sigma_{TOA} + (c\sigma_{TOA})^2 && (3.31)\\ &= d_{m,n} \pm 2c\sqrt{d_{m,n}}\sigma_{TOA} + c^2\sigma_{TOA}^2 && (3.32)\\ &= d_{m,n} \pm \sigma_{Dist}(d_{m,n}), && (3.33) \end{aligned}$$

where

$$\sigma_{Dist}(d_{m,n}) = 2c\sqrt{d_{m,n}}\sigma_{TOA} + c^2\sigma_{TOA}^2. \tag{3.34}$$

It is observed that each square distance measurement experiences a different uncertainty. However, there is no clear way to compute $\|\mathbf{N}_c\|_2$ from the measured data for each column. Therefore, a different model for considering the noise is needed to bound the norm in (3.29).

Consider that the measured squared distance $\hat{d}_{m,n}$ can be expanded as

$$\hat{d}_{m,n} = d_{m,n} + 2\sqrt{d_{m,n}}w_{m,n} + w_{m,n}^2, \tag{3.35}$$

where $w_{m,n}$ is the perturbation in the (m, n)-pair measurement. After the orthogonal projection is applied to a stacked version of (3.35), the following residual is obtained

$$\Pi_{\mathbf{R}}^{\perp}\hat{\mathbf{D}}_c = \Pi_{\mathbf{R}}^{\perp}\left[2\mathrm{diag}(\mathbf{w}_c)\mathbf{D}_c^{\circ\frac{1}{2}} + \mathrm{diag}(\mathbf{w}_c)\mathbf{w}_c\right] \in \mathbb{R}^{M\times 1}, \tag{3.36}$$

where $\mathbf{A}^{\circ p}$ denotes the p-th Hadamard power of the matrix $\mathbf{A}$ and diag($\mathbf{a}$) a diagonal matrix which non-zero entries are given by the vector $\mathbf{a}$.

Therefore, it is possible to provide a selection rule similar to (3.27) by upper bounding the square norm of the expression in (3.36). An appropriate upper bound for the residual norm can be given by

$$\begin{aligned}
\|\Pi_{\mathbf{R}}^{\perp}\hat{\mathbf{D}}_c\|_2^2 &= \|\Pi_{\mathbf{R}}^{\perp}[\mathrm{diag}(\mathbf{w}_c)(2\mathbf{D}_c^{\circ\frac{1}{2}} + \mathbf{w}_c)]\|_2^2 && (3.37)\\
&\leq \|\Pi_{\mathbf{R}}^{\perp}\|_2^2\|\mathrm{diag}(\mathbf{w}_c)\|_2^2\|2\mathbf{D}_c^{\circ\frac{1}{2}} + \mathbf{w}_c\|_2^2 && (3.38)\\
&\leq 4\sigma_{max}^2\bigg(\mathrm{diag}(\mathbf{w}_c)\bigg)\|\mathbf{D}_c^{\circ\frac{1}{2}} + 0.5\mathbf{w}_c\|_2^2 = \kappa_c, && (3.39)
\end{aligned}$$

where the fact that $\|\Pi_{\mathbf{R}}^{\perp}\| = 1$ has been used. Levering in (3.39), the subset of feasible combinations can be selected as

$$\mathcal{C} = \{c : f(c) \leq \kappa_c\}, \tag{3.40}$$

and an estimate of the distance matrix can be obtained using expression (3.28). Even though the bound in (3.39) always holds, it is not directly available from the measurements. In practice, only realizations of the measurement process are available, so in order to utilize the bound in (3.39) we introduce

$$\kappa_c^{(i)} = 4\gamma^i\sigma_{\mathbf{w}}^2\|\hat{\mathbf{D}}_c^{\circ\frac{1}{2}}\|_2^2, \;\; \gamma \geq 1, \tag{3.41}$$

as surrogate to provide a practical iterative threshold for the functional. In this expression $\sigma_{\mathbf{w}}^2$ and $\|\hat{\mathbf{D}}_c^{\circ\frac{1}{2}}\|_2^2$ denote the noise power and the norm squared from the Hadamard root of the c-th column of the measured distances matrix, respectively. The power of the noise can be safely considered known as it is assumed that the accuracy of the estimation method employed for the TOAs is known. For simplicity, it is assumed that all columns are subject to the same noise level σ_w. This assumption affects the performance of the bound as sources located at different positions have different accuracy levels. However, this can be seen as a reasonable assumption as the ordering of TOAs

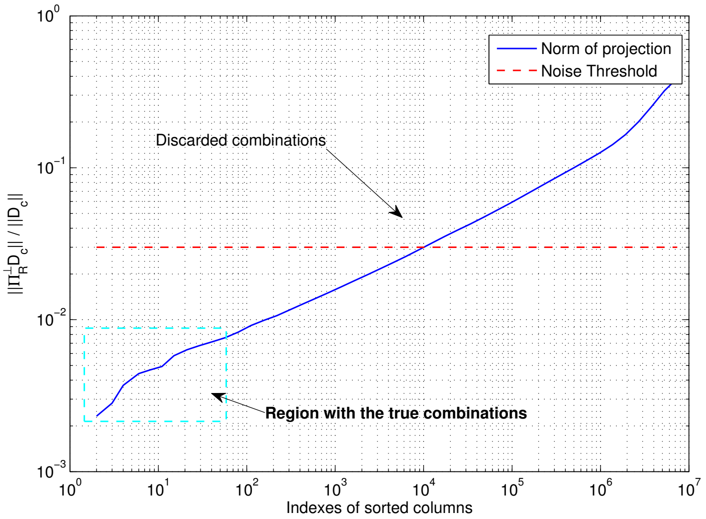


Figure 3.5: Illustration normalized norm of projection and noise threshold for $\sigma_w = 1.5$cm.

is unknown. In practice, it has been observed that $\kappa_c^{(0)}$ is enough for the method to deliver adequate results.

Finally, in contrast with the noise-free case, when measured data is available there is no guarantee that $|\mathcal{C}| = N$. In most of the cases, the threshold in (3.40) is overestimated (see Fig. 3.5). Despite this, the method is able to reduce the number of columns of the distance matrix, up to the noise accuracy, with a lower computational cost $\sim \mathcal{O}(M^2)$ in comparison with the $\sim \mathcal{O}(M^3)$ from the SVD decomposition for augmented EDMs.

3.3.2 Avoiding the Graph Problem

As discussed before, under real measurements the cardinality of $\mathcal{C}$ differs from the number of sources. Therefore, if the functional threshold is overestimated, further processing will be required to only select the appropriate columns. For this step two possible strategies can be applied: (i) the graph-based method discussed in the previous section, where the sorting problem is formulated as a maximum independent set problem, or (ii) a greedy approach based on the observation that two columns from $\mathbf{D}$ must not share elements in common and on the rank constraint for the augmented EDMs. As it is desired to avoid solving the NP-hard problem of listing all maximal independents sets, this subsection only focuses on the greedy solution of the problem.

After a filtered version of the distance matrix $\hat{\mathbf{D}} \in \mathbb{R}^{M\times|\mathcal{C}|}$, where $|\mathcal{C}| \ll N^M$ is obtained, it is possible to use the rank constraint of the augmented EDMs to further reduce the number of columns of $\hat{\mathbf{D}}$ as in the pre-filtering step of the previous section. This step excludes TOAs combinations that, approximately, violate the rank constraint without checking the rank constraint over all the set of N^M combinations. However, establishing an initial (or fixed) threshold ϵ is not as straightforward as for the subspace functional of (3.26). Therefore, an iterative approach is used to obtain the optimal ϵ starting at a low ϵ_0.

Combining two key observations, it is possible to develop a greedy strategy which

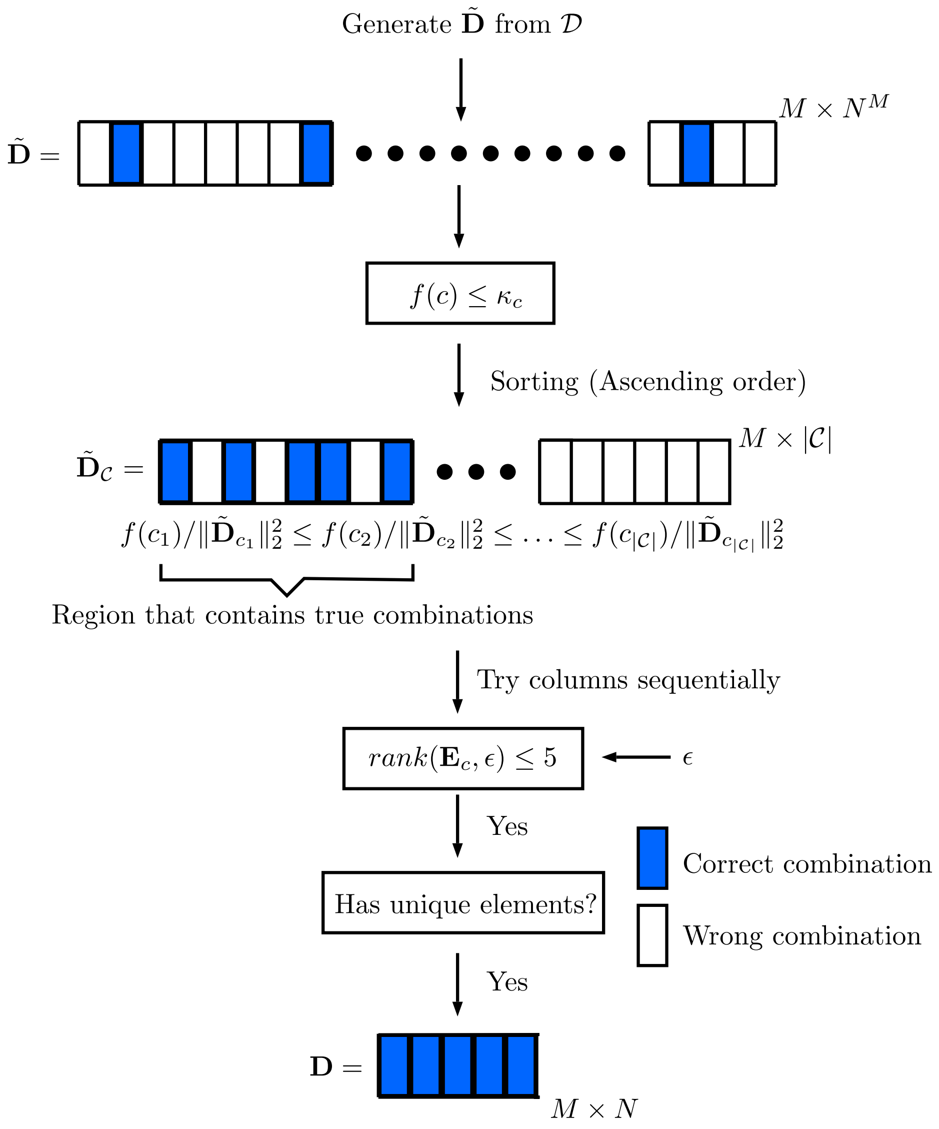

Generate $\tilde{\mathbf{D}}$ from $\mathcal{D}$
$\tilde{\mathbf{D}} =$
$M \times N^M$
$f(c) \leq \kappa_c$
Sorting (Ascending order)
$\tilde{\mathbf{D}}_{\mathcal{C}} =$
$M \times |\mathcal{C}|$
$f(c_1)/\|\tilde{\mathbf{D}}_{c_1}\|_2^2 \leq f(c_2)/\|\tilde{\mathbf{D}}_{c_2}\|_2^2 \leq \ldots \leq f(c_{|\mathcal{C}|})/\|\tilde{\mathbf{D}}_{c_{|\mathcal{C}|}}\|_2^2$
Region that contains true combinations
Try columns sequentially
$rank(\mathbf{E}_c, \epsilon) \leq 5$
ϵ
Yes
Has unique elements?
Yes
Correct combination
Wrong combination
$\mathbf{D} =$
$M \times N$

Figure 3.6: General flow of the greedy strategy for sorting the acoustic echos

overcomes the need of solving the maximum independent set problem. Firstly, the fact that it is very unlikely for the columns of $\hat{\mathbf{D}}$ to have elements in common (starting point for the maximum independent set problem of the graph approach) is used. Secondly, by using the functional $f(c)$ for sorting the columns of $\tilde{\mathbf{D}}$, it is seen that the columns with the lowest normalized functional value, meeting the ϵ-rank constraint, has to be part of the true distance matrix. The rank constraints alleviates the problem shown in Fig. 3.4 where the sharp knee in the graph of the sorted values of $f(c)$ starts to become smoother by the presence of noise. The algorithm combining these two key observations is presented in Algorithm 1 and the general flow of the method in Fig. 3.6. In Algorithm 1, the parameter $\eta > 1$ controls the growth of the rank constraint. This allows the method to only introduce the best ranked columns to the solution.

Furthermore, in Table 3.1 a comparison between the graph-based and the subspace-based methods, in terms of of the computational complexity for each step, is shown.

This summary is not, by any means, strict. It is intended to provide a sense of the complexity for each of the main steps in the methods, in order to highlight the reduction achieved by using subspace filtering to reduce the set of feasible combinations.

Algorithm 1 Subspace-based Greedy TOA Sorting

```
Input: $\mathcal{D}$, $\Pi_{\mathbf{R}}^{\perp}$, $\mathbf{E}$, $\epsilon_0$, $N$, $\boldsymbol{\sigma}_{\mathbf{w}}$
Output: $\mathbf{D}$
    Initialization: Generate $\tilde{\mathbf{D}}$ and $\boldsymbol{\kappa}^{(0)}$, $\mathbf{D} = \{\}$, $\epsilon = \epsilon_0$
$\mathcal{C} = \{c : f(c) \leq \boldsymbol{\kappa}_c^{(0)}\}$
$\mathcal{C}_s = \mathrm{sort}(\mathcal{C}, f(c)/\|\tilde{\mathbf{D}}_c\|_2^2, ”\mathit{ascending}”)$
$\hat{\mathbf{D}} = \tilde{\mathbf{D}}_{\mathcal{C}_s}$
while numCols($\mathbf{D}$) $< N$ do
    for $c = 1$ to $|\mathcal{C}_s|$ do
        $\tilde{\mathbf{E}} = \begin{bmatrix} \mathbf{E} & \hat{\mathbf{D}}_c \\ \hat{\mathbf{D}}_c^T & 0 \end{bmatrix}$
        if rank($\tilde{\mathbf{E}}, \epsilon$) $\leq 5$ and $\hat{\mathbf{D}}_c \cap \mathbf{D} == \emptyset$ then
            $\mathbf{D} = [\mathbf{D}, \hat{\mathbf{D}}_c]$
        end if
    end for
    if numCols($\mathbf{D}$) $< N$ then
        $\epsilon = \eta\epsilon$
    end if
end while
```

It is important to remark that contrary to the graph-based method, where more than one maximum independent set can be found in the graph, this greedy alternative provides a unique solution. The unique solution allows the method to sort the TOAs even when the constraint imposed by Polleyfey's method is not met.

Finally, after the matrix $\mathbf{D}$ is estimated by the greedy approach, the least squares solution for the estimates of the source locations, for $M \geq 5$, can be directly obtained by

$$\hat{\mathbf{S}} = (\mathbf{R}^T)^{\dagger}\mathbf{D}. \tag{3.42}$$

Notice that if distances matrices estimates from Q acoustic sources are available, i.e.,

$$\mathbf{D}_{\text{Tot}} = \big[\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_Q\big] = \mathbf{R}^T\big[\mathbf{S}_1, \ldots, \mathbf{S}_Q\big], \tag{3.43}$$

a combination of the Polleyfey's method, using the SVD of $\mathbf{D}_{\text{Tot}}$, and Procrustes analysis could be performed to estimate the image source positions instead of using (3.42). This approach could lead to better reconstruction results for cases in which the pseudo-inverse of $\mathbf{R}^T$ is not well conditioned. As commented in the graph-based approach, after the (image) sources locations in $\hat{\mathbf{S}}$ are obtained, the room boundaries can be obtained by straightforward geometrical methods.

Graph-based		**Subspace-based (Greedy)**	
Step	**Complexity**	**Step**	**Complexity**
Rank filter	$N^M\mathcal{O}((M+1)^3)$	Subspace filter	$N^M\mathcal{O}(M^2)$
Maximum Ind. Sets	$\mathcal{O}(2^{0.276\lvert\mathcal{C}_\epsilon\rvert})$	Rank filter	$\lvert\mathcal{C}\rvert\mathcal{O}((M+1)^3)$
Pollefey's + Procrustes	$\left[\prod_{i=1}^{Q} \lvert\mathcal{S}_{max}^{G_i}\rvert\right]\mathcal{O}(49MQ^2)$	Least Squares	$\mathcal{O}(NM^2)$

Table 3.1: Complexity comparison of the steps of the graph-based and the greedy alternative

3.4 Experimental Results

In this section, results from a set of simulations are presented to evaluate the performance of the greedy approach proposed in this chapter with respect to number of sources, number of microphones, and number of measurements. In addition, the current state-of-the-art method, based on graph theory, is compared with the proposed greedy alternative.

For each evaluated parameter a set of 500 Monte Carlo simulations are performed using synthetic data created by placing sources and microphones randomly in a room. The distances between the microphones and the image sources are computed and perturbed with white Gaussian noise with power σ^2, i.e.,$\sim \mathcal{N}(0,\sigma^2)$, to simulate the uncertainties in the TOA estimates. The experiments are run in Matlab on a Macbook Air (Mid 2013) 1.7 GHz Inter Core i7 processor.

The room used in these experiments has a constant volume of $280m^3$ with dimensions $8m \times 6m \times 5m$. Furthermore, the RMSE used to quantify the methods performance is the expectation of the square root of the error squared between estimated and true $3D$ room vertices. That is,

$$\text{RMSE}(\hat{\boldsymbol{\theta}}) \triangleq \sqrt{E\left[\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2\right]} \tag{3.44}$$

where $\boldsymbol{\theta} \in \mathbb{R}^8$ represents the $3D$ room vertices and $\hat{\boldsymbol{\theta}} \in \mathbb{R}^8$ the estimates of the vertices given by the method.

3.4.1 Number of Sources

As discussed in the previous section, the uncertainty in the TOA estimates, i.e., σ_{TOA}, imposes an intrinsic limitation to the estimation of the room reconstruction. This leads to degradation in the method performance.

One alternative to diminish the effect of the TOAs uncertainty is to rely on more than one source. By averaging the different vertices estimates, obtained by using L sources, it is possible to increase the estimate accuracy. For these simulations, we consider the case of $M = 7$ microphones distributed randomly in our shoe-box shaped room. The vertices estimates $\hat{\boldsymbol{\theta}}_l$ are obtained from the image sources locations estimates, for each of the L sources, and then averaged. The RMSE is then defined as

$$\text{RMSE}(\hat{\theta}) \triangleq \sqrt{E\left[\left\|\frac{1}{L}\sum_{l=1}^{L}\hat{\boldsymbol{\theta}}_l - \boldsymbol{\theta}\right\|_2^2\right]} \tag{3.45}$$

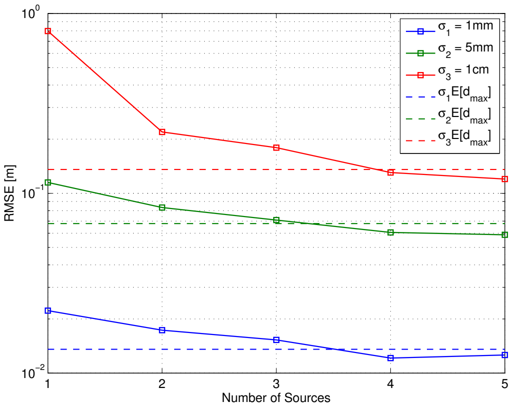


Figure 3.7: RMSE of room reconstruction as function of number of sources and TOA uncertainty

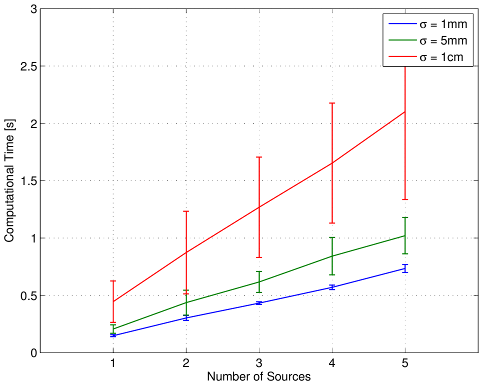


Figure 3.8: Computation time for different TOA uncertainties and number of sources

By averaging the vertices estimates, across the available sources, the RMSE can be reduced. This result is shown in Fig. 3.7. Furthermore, notice how for higher $\sigma = c\sigma_{TOA}$ the RMSE of the estimates also increases. As the number of estimates used to averaging increases, the RMSE start reaching the maximum position uncertainty, i.e., $\sigma_{Dist} \approx c\sigma_{TOA}d_{max}$. This is the approximate uncertainty that the farthest image source has. Fig. 3.8 shows that the reduction in RMSE requires a linear increase in computational time. The increase in computation time due to TOA estimates uncertainty is explained by the adaptive threshold ϵ of the rank-constraint. Larger uncertainties require more iterations to find an appropriate threshold.

3.4.2 Number of Receivers

Increasing the number of receivers does not reduce the uncertainty due to the TOA estimates, but provides a more selective kernel for the complementary projection matrix

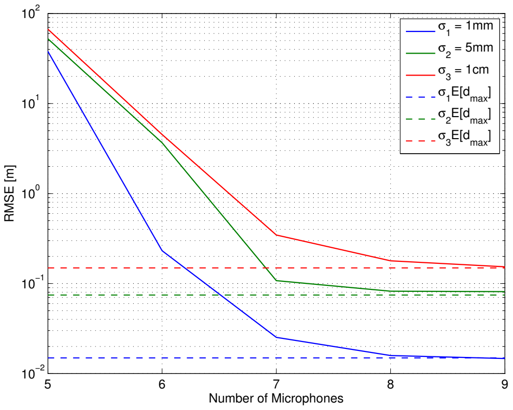


Figure 3.9: RMSE of room reconstruction as function of number of microphones and TOA uncertainty

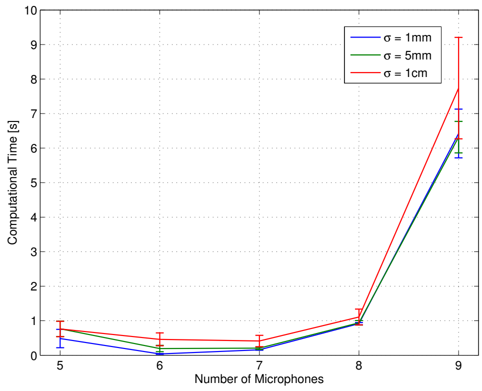


Figure 3.10: Computation time for different TOA uncertainties and number of microphones

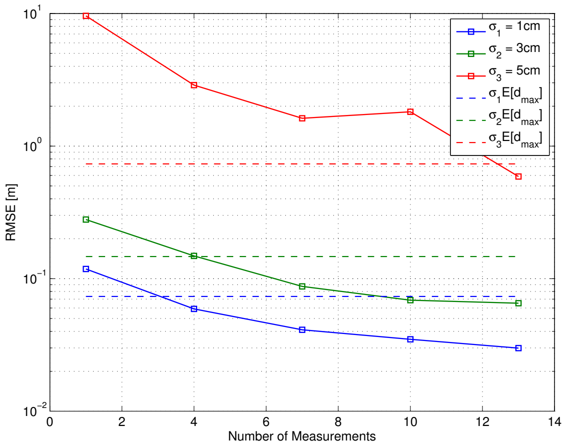


Figure 3.11: RMSE of room reconstruction as function of number of measurements

reducing the error in the estimation. Due to the constraint in the rank of our matrices, the minimum number of receivers that can be employed is 5. However, as seen in Fig. 3.9, the $ker(\mathbf{R})$ for $M = 5$ does not provide a very selective subspace, leading to higher RMSE. As the number of receivers increases the $ker(\mathbf{R})$ becomes better defined and filters out many more vectors that could cause ambiguities in the solution. However, by increasing the number of microphones the number of combinations to test increases exponentially and so does the computation time of the method as shown in Fig. 3.10.

The counter intuitive behavior in Fig. 3.10, where convex curves are seen is explained by the selectivity of the kernel. As the kernel in $M = 5$ is not selective at all, many more combinations have to be checked compared to the cases $M = 6$ and $M = 7$ where more selective kernels are available, increasing the time that the method consumes.

3.4.3 Number of Measurements

Instead of increasing the number of receivers or sources to obtain more measurements, another alternative is to perform multiple estimations of the TOAs. This is equivalent to estimate the RIRs, at each microphone, several times. If the used RIR estimator is unbiased, averaging over several realizations will provide better estimates of the TOAs. Contrary to the graph-based approach, where more than one source is needed, the greedy strategy can rely on several measurements of the same distribution of microphones and source to improve its estimation performance. Fig. 3.11 shows how by increasing the number of times that the TOAs are estimated the RMSE of the reconstruction diminish. In this case, the only increase in computation time will be due to the multiple estimates of the TOAs.

3.4.4 Comparison Greedy Approach vs (modified) Graph-based method

In this part we compare the performance of the proposed greedy alternative and the graph-based method. Since the graph-based approach requires more than one source to disambiguate its results, a version of it using an oracle is employed in the experiments. The oracle has as output the maximum independent set with the minimum estimation

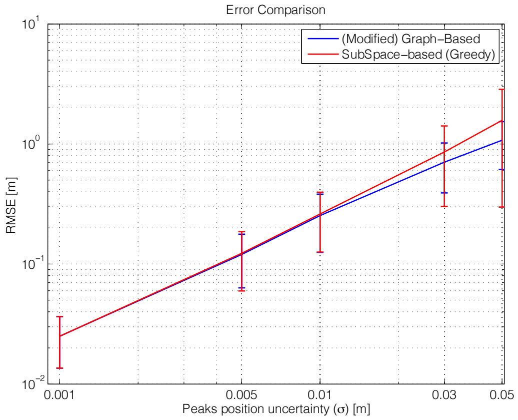


Figure 3.12: Error comparison between the proposed greedy strategy and the modified graph-based method

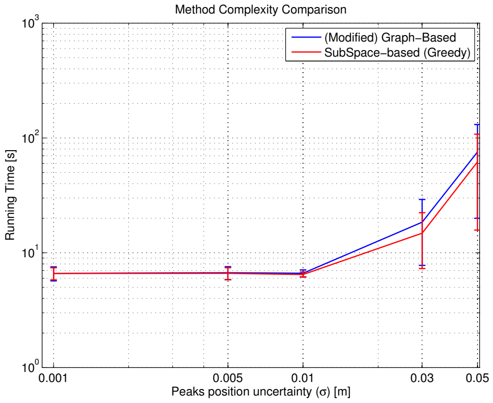


Figure 3.13: Computation time comparison between proposed method and the modified graph-based approach

error with respect to the true distance matrix. This allows the graph-based approach to deliver a result using a single source, always returning the best candidate in terms of reconstruction error. In addition, as the graph-based approach is unable to provide results in reasonable time for $M > 7$, the method has been modified by including the subspace filtering to significantly reduce its running time. The configuration of the experiments, for both methods, includes $M = 9$ microphones and $N = 1$ source.

In Fig. 3.12 and Fig. 3.13 the error and the computation time of both methods are shown for different TOA uncertainties. For low uncertainties in the TOA estimates, both methods present the same performance. However, at higher uncertainties, the greedy approach presents worse performance than the modified graph-based method. This difference in performance can be explained by the rejection rate of the methods. This rate is the percentage of rejected datasets, which can not be used to produce source estimates by the methods. In Fig. 3.14 the rejection rate for both methods is compared. Notice how the graph-based approach has a very high rejection percentage when the TOA uncertainty is high. This occurs in situations in which no maximum set are found or, due to time constraints, when the process halts as result of the large number of nodes in the graph. In the latter case the solution could have been found, however the time needed to find the solution of the NP-hard problem renders the approach infeasible in acceptable time. On the contrary, the greedy alternative only rejects a dataset if the positions of the microphones are collocated. Note that in this case, the graph-based approach also rejects the dataset.

Notice that in Fig. 3.13 the modified graph-based approach shows a similar computation complexity as the greedy strategy. This is explained by the subspace pre-filtering stage. As the subspace filtering removes a large quantity of infeasible combinations, most of the computational load is due to this stage. If the same experiment were to be tested without the subspace filtering, evaluating the 6^9 combinations using the graph-based method becomes infeasible as reported in [24].

Finally, to illustrate the trade off between computational time and performance when using the suboptimal greedy strategy, a comparison between the graph-based

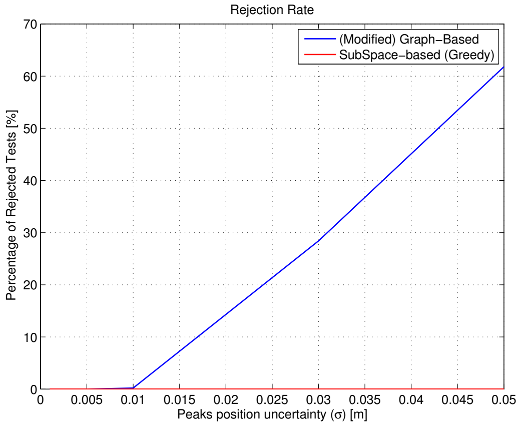


Figure 3.14: RMSE of room reconstruction as function of number of measurements

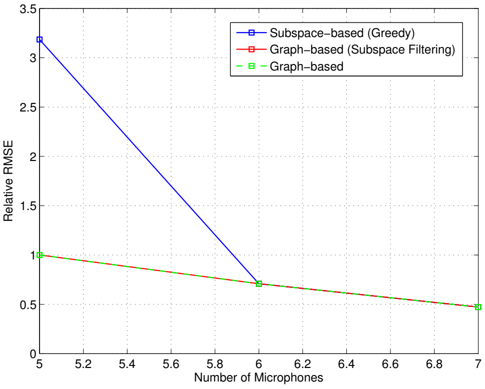


Figure 3.15: Relative error comparison between the proposed greedy strategy and the graph-based methods

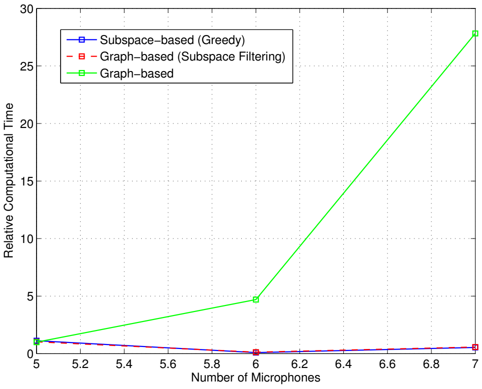


Figure 3.16: Computation time comparison between proposed method and the graph-based methods

method, the modified graph-based method (subspace filtering added) and the proposed greedy alternative is presented. In this comparison, the limiting case of small errors in the RIRs peaks, i.e., $\sigma_{RIR} = 0.5$mm, is considered. The comparison, in terms of relative computational complexity and relative RMSE is shown in Fig. 3.15 and Fig. 3.16.

The relative computational time and relative RMSE are computed by normalizing the performance results of each method to the respective performance of the graph-based method when $M = 5$ microphones are employed. The baseline RMSEs, provided by the (original) graph-based are given in Table 3.2.

Microphones	5	6	7
RMSE	171.6mm	0.5mm	0.2mm

Table 3.2: Graph-based RMSEs used as baseline values

From Fig. 3.15, the effect caused by the selectivity of the kernel for $M = 5$ and the usage of the pseudo inverse of $\mathbf{R}^T$, as described in previous subsections, is seen. This

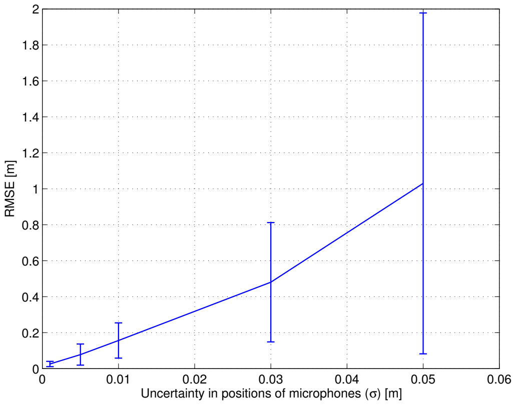


Figure 3.17: RMSE of the vertices vector for different uncertainties in the microphones

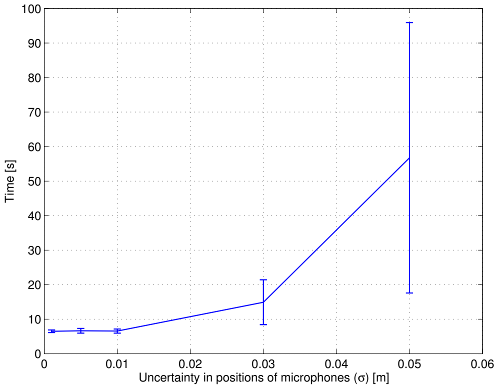


Figure 3.18: Computation time for different uncertainties in the microphones

issue makes the error of the greedy alternative three times larger than the graph-based method. However, as the number of microphones increases, the RMSE of the greedy alternative reaches the RMSE of the graph-based methods. Notice that adding the subspace filtering does not affect the RMSE of the graph-based method. In addition, in Fig. 3.16 the reduction in the computational time when the subspace filtering is applied is clearly seen. While the computational time of the greedy approach and the modified graph-based method remains almost unchanged, the computational time of the original graph-based explodes in an exponential fashion. These results show two main things: (i) the greedy alternative does reach the performance of the graph-based method for increasing number of microphones at a reduced computational cost, and (ii) if the the graph-based approach is preferred for any particular reason, the inclusion of the subspace filter reduces the computational time of the original method while preserving its optimality. Further comparisons considering $M > 7$ are not performed as the (original) graph-based, most of the times, becomes intractable.

3.4.5 Uncertainty in the Microphone Positions

The weakest point of the proposed method is its dependency on the precise locations of the microphones. The whole approach is based on the premise that these positions are known with high accuracy. In this section, it is shown how uncertainties in the positions of the microphones affect the estimation of the room vertices. A set of 500 Monte Carlo simulations were performed using $M = 9$ microphones and a single source. An uncertainty of $\sigma_{RIR} = 2$mm in the peaks of the RIRs is assumed. The reconstruction is performed using the pseudo inverse of the microphones positions matrix $\tilde{\mathbf{R}}^T$ built with the noisy microphones locations.

From Fig. 3.17 it is seen that the performance of the greedy strategy degrades fast as the uncertainty in the positions of the microphones increases. As the estimated subspace is not representative for the distance matrix, the assumption that the first columns in the sorted distance matrix belong to the true combinations of echoes does not hold anymore. This issue heavily affects the estimation of the room vertices even at low uncertainties in the positions of the microphones. Fig. 3.18 shows the increased in computational time due to the noise in the TOA estimates and positions of the microphones.

To alleviate this problem, it is possible to rely on multiple measurements using more than one acoustic source and the combination of Pollefey's method with Procrustes analysis. However, a straightforward implementation of this approach does not guarantee the proper reconstruction of the room vertices. This is due to the suboptimality of the greedy strategy, i.e., if many of the combinations of echoes are wrongly estimated, the Pollefey's method will not deliver adequate results. Despite this issue, it is possible to use a technique based on bagging [44][6], i.e., bootstrap aggregating, to improve the stability and accuracy of the reconstruction of the vertices. This involves a sampling process to diversify (increase) our data set to perform the estimation. That is, by randomly sampling the columns from the estimated distance matrices, multiple estimates for the microphone positions using the Pollefey's method can be created. By selecting the best position estimates, when the error is considered as the difference with respect to the noisy locations of the microphone after Procrustes analysis, the image sources, and hence the room vertices can be estimated. This process is summarized in Algorithm 2.

Algorithm 2 Random sampling strategy for improving stability of room vertices estimates

```
Input: ${\mathbf{D}_i}_{i=1}^{Q}$, $\mathbf{micPos}$, maxTrials, nSrcs2Use
Output: $\mathbf{estVertices}$
    Initialization: Generate $\mathbf{D} = [\mathbf{D}_1, \ldots, \mathbf{D}_Q]$
[$\sim$, nImgSrc] = size($\mathbf{D}$)
for $i = 1$ to maxTrials do
    $\mathbf{v}$ = randperm(1 : nImgSrc)
    $\mathbf{v} = \mathbf{v}$[1 : nSrcs2Use]
    [$\mathbf{R}, \mathbf{S}$] = pollefeys( $\mathbf{D}$[:, $\mathbf{v}$] )
    [$\mathbf{T}$, fit] = estimateRigidTransform($\mathbf{micPos}, \mathbf{R}$)
    $\mathbf{estMicPos}$ = applyTransform($\mathbf{T}, \mathbf{R}$)
    micErr($i$) = $\|\mathbf{micPos} - \mathbf{estMicPos}\|_2$
    imSrcList{$i$} = $\mathbf{S}$
    listT{$i$} = $\mathbf{T}$
    listMics{$i$} = $\mathbf{estMicPos}$
end for
bestC = find($\mathbf{micErr}$ == min($\mathbf{micErr}$))
$\mathbf{bestMicPos}$ = listMics{bestC}
$\mathbf{estImSrc}$ = applyTransform(listT{bestC}, imSrcList{bestC})
$\mathbf{estVertices}$ = findVertices($\mathbf{bestMicPos}, \mathbf{estImSrc}, \mathbf{D}$)
```

Results from applying the proposed alternative based on random sampling for $Q = 4$ acoustic sources and $M = 9$ microphones are shown in Fig. 3.19 and Fig. 3.21. In this instance, it is assumed that the peaks in the RIRs are estimated with an accuracy of $\sigma_{RIR} = 1$cm. From these figures, it is possible to observe how the proposed approach considerably reduces the error in the reconstruction. By using more than one source and stabilizing the estimation using random sampling it is possible to achieve estimates

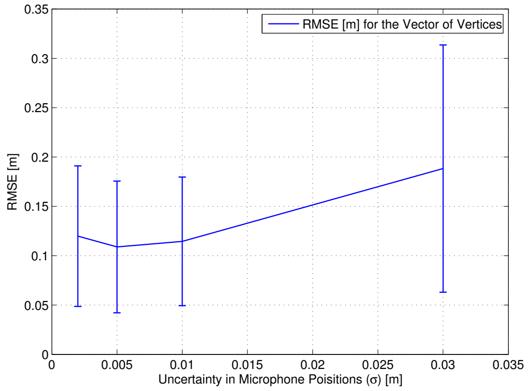


Figure 3.19: RMSE of the vertices vector for different uncertainties in the microphones using random sampling

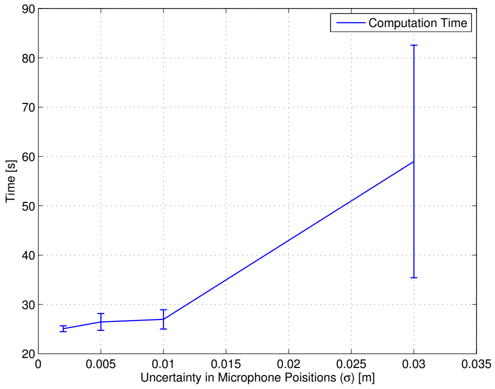


Figure 3.20: Computational time for different uncertainties in the microphones using random sampling

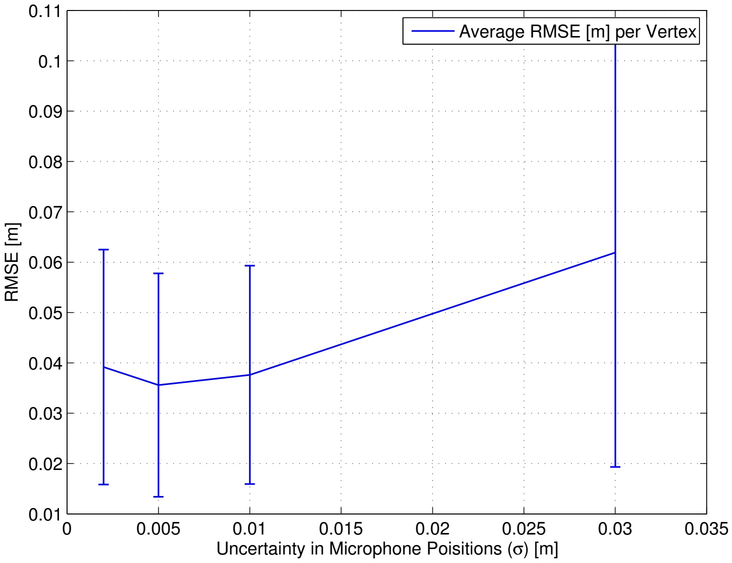


Figure 3.21: Average RMSE per vertex of the shoe-box shaped room when random sampling is used to estimate them

for the room vertices with an average error below 10cm for different uncertainties in the positions of the microphones. Notice that this method reduces the effect of both RIRs peaks and microphones positions uncertainties. The computational time increases as the uncertainty in the locations of the microphone increases. These results show that even if the microphone positions are not known with high accuracy, the greedy strategy can still be used to estimate the room vertices. An example for the reconstruction of the 3D room vertices is shown in Fig. 3.22.

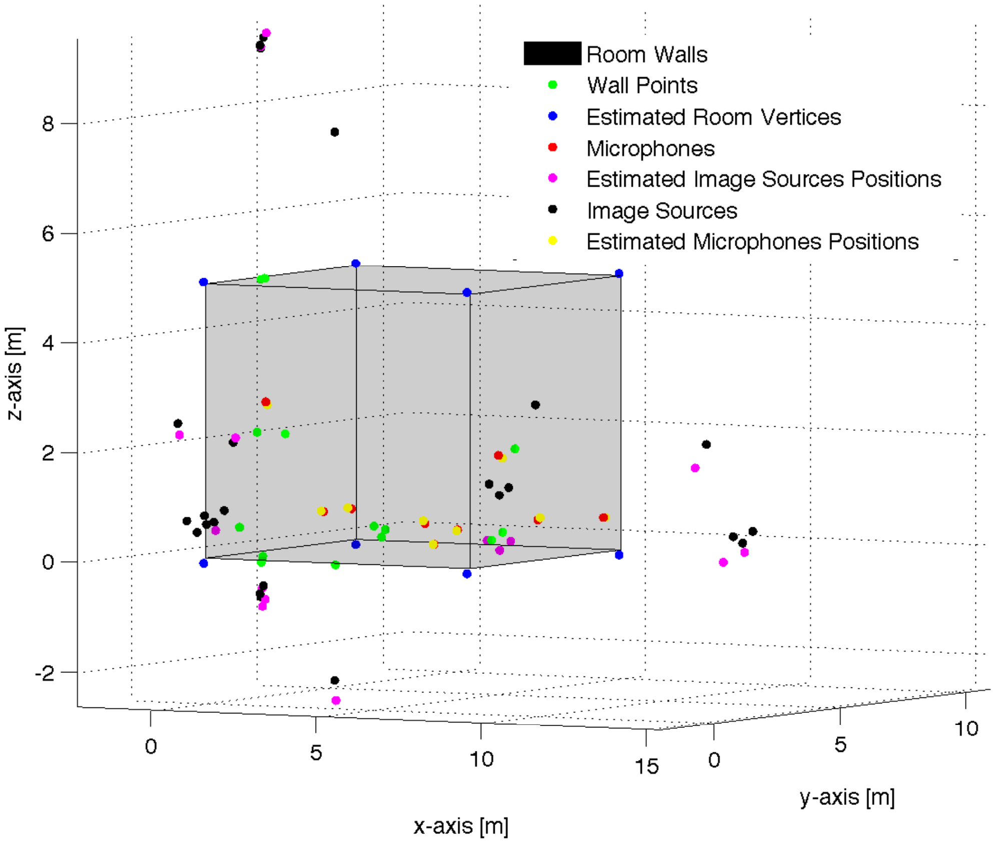


Figure 3.22: Example of a noisy reconstruction for the room vertices using the proposed method

3.5 Discussion

In this chapter a greedy alternative for the problem of acoustic echo sorting problem was presented. The method was motivated by the rank-5 factorization of the distance matrix $\mathbf{D}$. By exploiting the kernel of the matrix $\mathbf{R}$ positions of the microphones a strategy to identify columns of $\mathbf{D}$ was devised. It was shown that by using a greedy strategy, based on the uniqueness of columns of $\mathbf{D}$ and the rank constraint of EDMs, it is possible to identify columns of $\mathbf{D}$ even under the presence of noise. The greedy strategy was shown to outperform, in terms of computational complexity, the current state-of-the-art method based on a graph problem without compromising accuracy. In addition, the proposed method only requires one source, and its performance was shown to be bounded by the accuracy of the TOA estimates. In case of large uncertainties in both RIRs and microphone locations, a method based on random sampling was introduced to improve the stability of the estimation. This method was able to produce estimates with an average error smaller than 10cm by using $Q = 4$ sources. Finally, it should be noted that the subspace filtering, introduced in this thesis, can be used to enhance the graph-based method in terms of speed. However, even though these methods are able to estimate the room walls from a set of peaks in the RIRs, selecting these peaks can be a challenging task. This step can be even harder than the sorting of the peaks. Motivated by this issue, the next chapter discusses a general method, based on estimation theory, to estimate the image sources positions from microphone measurements, not necessary from measured RIRs. This approach solves both problems of finding and sorting the peaks only requiring the relative positions of the microphones and a controlled source.

Wideband CLEAN/RELAX for Source Localization 4

In this chapter, an alternative approach, spanning from estimation theory, is presented to find the positions of the first order reflections in a room. This problem is solved in an iterative manner by using a nonlinear least square (NLS) estimator, which is a surrogate for the maximum likelihood estimator (MLE) under white Gaussian noise. Even though this work is focused in the case of shoe-box shaped rooms, the proposed approach is applicable for arbitrary rooms shapes with flat walls.

4.1 Why is a different approach needed?

In the previous chapter it was shown that it is possible to find the first-order reflections, and estimate the room geometry, by means of an efficient technique based on the properties of the distance matrix $\mathbf{D}$. However, the methods were devised considering that the assumptions (*A.1*)-*(A.5)* always hold. These assumptions, although plausible in some circumstances, do not hold in many practical situations. Let us consider *(A.1)* for example. In this case, it is assumed that it is possible to *easily* find the peaks in a RIR which corresponds to the first-order reflections. However, in most measured RIRs this can be a challenging task. Not only the reverberation in the room, which partially buries the weaker reflections in the RIRs, affects the peaks picking task. Effects from the loudspeaker transfer function and directivity also hinder the ability to properly find the peaks in the RIRs. Inevitably, the loudspeaker colors and shapes the original RIR by mixing and attenuating several peaks. This makes harder to detect which peaks belong to the walls reflections. To illustrate this, a typical impulse response for a loudspeaker is shown in Fig. 4.1. In Fig. 4.2 the frequency response of the loudspeaker is shown, which resembles a band-pass filter for signals within the audible frequency band.

It can be argued that if the loudspeaker impulse response is known beforehand, as in the case of loudspeaker manufacturers, a deconvolution process can be carried out in order to obtain the original RIR. However, besides the fact that the convolution process losses information, i.e., depending on the impulse response the original signal could not be retrieved completely, the loudspeaker is not an omnidirectional source. Hence, reflections originated from different angles with respect the loudspeaker suffers different attenuation and/or changes of phase that are difficult to account for in practice. This poses a challenging problem for correctly selecting the peaks that belongs to the first-order reflections. As a result, the methods for sorting echoes need to consider as many peaks as possible. This explodes the computation time required for finding the correct combination of echoes. A comparison between an (ideal) sparse RIR and an actual measured RIR is shown in Fig. 4.3 and Fig. 4.4.

Besides the complications that a real loudspeaker could add to the problem of finding

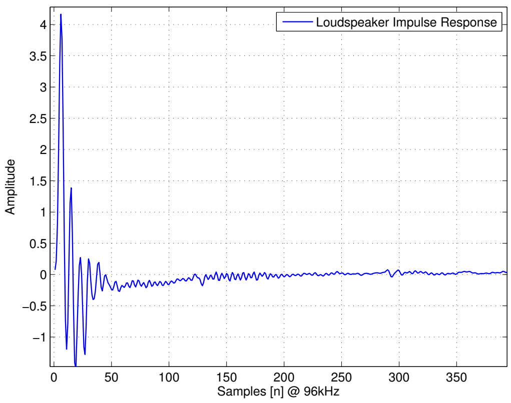


Figure 4.1: Example of loudspeaker impulse response measured at 96kHz

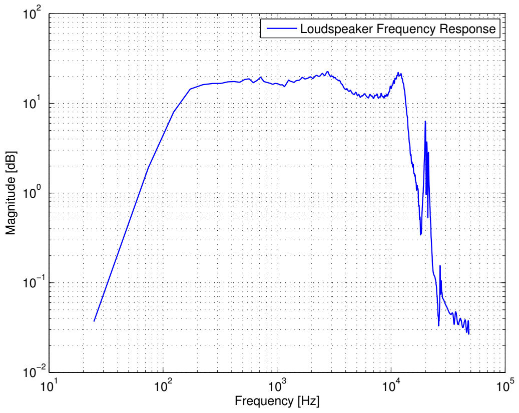


Figure 4.2: Frequency response of measured loudspeaker

first-order reflections, it is necessary to consider some other practical matters. In the previous chapter, it is assumed that the TOAs are available through measured RIRs ($A.2$). However, most of the products on the market, aimed for reproduction purposes, do not measure the RIRs, and the ones capable of measuring the RIRs require probe signals unpleasant to the occupants in the room. Hence, it is of interest to develop methods capable to perform the estimation of the first-order reflections using amenable sounds for the people inside the room.

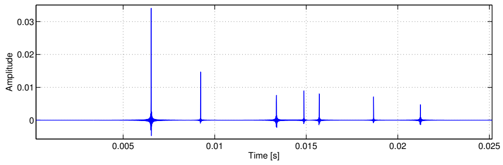


Figure 4.3: Sparse (ideal) RIR

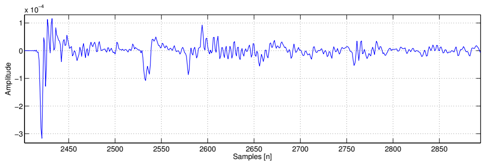


Figure 4.4: Measured RIR

Therefore, this chapter is aimed to develop general methods, based on estimation theory, to find the first-order reflections. The approaches use raw measurements from the available microphones in the room to iteratively search for the first-order reflections. In principle, these methods can employ an arbitrary signal for finding the locations of the reflections. In addition, knowledge of the loudspeaker transfer function can be added in order to cope with effects induced by the drivers in the transmitted signal. Furthermore, due to our interest in common audio reproduction systems, setups as the one depicted in Fig. 4.5 are our main focus. In this scenario, a set of loudspeakers are equipped with microphone arrays. Particularly, a uniform circular arrays of 6cm radius, consisting of $M = 3$ microphones, is considered for each of the loudspeakers.

In the following section, the data modal used to develop the iterative method is introduced.

4.2 Data Model

Consider a set $\mathcal{M}$ of M microphones randomly distributed in a shoe-box shaped room. For a given excitation signal $s(t)$, the data acquired by the m-th microphone is given by

$$x_m(t) = \sum_{i=0}^{|\mathcal{I}|} \gamma_i \beta_{mi} s(t - \tau_{mi}) + w_m(t), \tag{4.1}$$

where $\beta_{mi} \in \mathbb{R}$ and $\tau_{mi} \in \mathbb{R}$ represent the attenuation and propagation delay at the m-th microphone related to the i-th (image) source, and $\gamma_i \in \mathbb{R}$ is the attenuation due to the reflective surface. The set $\mathcal{I}$, as defined in (2.9), is the set of reflections in the room considered in the model and the case $i = 0$ is reserved for the contribution of the direct path, i.e. $\gamma_0 = 1$. In this model, the noise $w_m(t)\,\forall\, m$ is considered to be white Gaussian noise with power σ_w^2.

Considering spherical isotropic radiators as in Chapter 2, the attenuation and delay at the m-th microphone with respect the i-th source are given by

$$\beta_{mi} = \frac{1}{4\pi \|\mathbf{r}_m - \mathbf{s}_i\|_2}, \quad \tau_{mi} = \frac{\|\mathbf{r}_m - \mathbf{s}_i\|_2}{c}, \tag{4.2}$$

where $\mathbf{r}_m \in \mathbb{R}^3$ and $\mathbf{s}_i \in \mathbb{R}^3$ are the spatial positions of the m-th microphone and the i-th source, and c is the speed of sound. The model in (4.1) is obtained from the convolution of the RIR in (2.9) and an excitation signal $s(t)$.

Furthermore, as this work is focused in audio reproduction setups, where there is control over the excitation signal $s(t)$, it is assumed that $s(t)$ is a zero-mean, real-valued and periodic signal. That is, the sampled transmitted signal allows an expansion in harmonic functions given by [26]

$$s[n] = \sum_{k=1}^{Q} A_k \cos(\omega_0 k n + \phi_k) = \sum_{k=-Q}^{Q} \alpha_k \exp(j\omega_0 k n), \tag{4.3}$$

for $n = 0, 1, \ldots, N-1$ where $A_k > 0, \phi_k \in [-\pi, \pi)$, $\omega_0 \in (0, \pi/Q)$, and $\alpha_k = \alpha_k^* = A_k \exp(j\phi_k)/2$ are the amplitude, phase, fundamental frequency, and complex amplitude, respectively. As it is considered that the signal is zero-mean, there is no DC component in the expansion (4.3), i.e., $\alpha_0 = 0$.

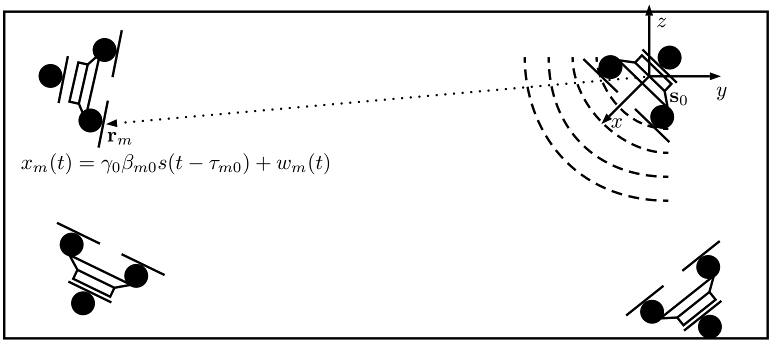


Figure 4.5: Sound reproduction setup illustrating the practicalities of the arrangement

Under this model, if the source signal is delayed by a delay $\eta_{mi} = f_s \tau_{mi}$, where f_s denotes the sampling frequency, the following is obtained

$$s[\eta_{mi}] \triangleq s[n - \eta_{mi}] = \sum_{k=-Q}^{Q} \alpha_k \mathrm{exp}(j\omega_0 kn)\mathrm{exp}(-j\omega_0 \eta_{mi}). \tag{4.4}$$

Using matrix-vector notation, it is possible to rewrite a stacked version of (4.4), for $n = 0, 1, \ldots, N$, as

$$\mathbf{s}[\eta_{mi}] = \mathbf{Z}(\omega_0)\mathbf{D}(\eta_{mi})\boldsymbol{\alpha}, \tag{4.5}$$

where the following has been defined

$$\begin{aligned}
\mathbf{z}(\omega) &\triangleq [1\ \mathrm{exp}(j\omega)\ \cdots\ \mathrm{exp}(j\omega(N-1))]^T, && (4.6)\\
\mathbf{Z}(\omega_0) &\triangleq [\mathbf{z}(-Q\omega_0)\ \cdots\ \mathbf{z}(-\omega_0)\ \mathbf{z}(\omega_0)\ \cdots\ \mathbf{z}(Q\omega_0)], && (4.7)\\
\mathbf{D}(\eta_{mi}) &\triangleq \begin{array}{l}\mathrm{diag}(\mathrm{exp}(jQ\omega_0\eta_{mi}), \ldots, \mathrm{exp}(j\omega_0\eta_{mi}),\\ \qquad\qquad \mathrm{exp}(-j\omega_0\eta_{mi}), \ldots, \mathrm{exp}(-jQ\omega_0\eta_{mi}))\end{array} && (4.8)\\
\boldsymbol{\alpha} &\triangleq [\alpha_{-Q}\ \cdots\ \alpha_{-1}\ \alpha_1\ \cdots\ \alpha_Q]. && (4.9)
\end{aligned}$$

If N samples for each microphone are recorded, and M data vectors stacked, the recorded sampled data can be expressed as

$$\begin{aligned}
\mathbf{x} &= [\mathbf{x}_1^T\ \mathbf{x}_2^T\ \cdots\ \mathbf{x}_M^T]^T && (4.10)\\
&= \sum_{i=0}^{|\mathcal{I}|} \gamma_i \mathbf{H}(\boldsymbol{\eta}_i, \boldsymbol{\beta}_i)\boldsymbol{\alpha} + \mathbf{w} \in \mathbb{R}^{MN\times 1}, && (4.11)
\end{aligned}$$

with $\mathbf{H}(\boldsymbol{\eta}_i, \boldsymbol{\beta}_i) = [\mathbf{H}_{1i}^H\ \mathbf{H}_{2i}^H\ \cdots\ \mathbf{H}_{Mi}^H]^H$, where

$$\mathbf{H}_{mi} = \beta_{mi}\mathbf{Z}(\omega_0)\mathbf{D}(\eta_{mi}). \tag{4.12}$$

Finally, as both β_{mi} and τ_{mi} are dependent on the position $\mathbf{s}_i$ of the i-th source, the dependency of $\mathbf{H}$ can be expressed in terms of the sources positions instead. Thus, the model in (4.11) can be rewritten as

$$\mathbf{x} = \sum_{i=0}^{|\mathcal{I}|} \gamma_i \mathbf{H}(\mathbf{s}_i)\boldsymbol{\alpha} + \mathbf{w}. \tag{4.13}$$

In the next sections, the model in (4.13) is used to estimate the positions of the image sources.

4.3 Localization of First Order Reflections

To accurately estimate the positions of the first order reflections, the wideband signal parameter estimation problem of the previous section needs to be solved. That is, from the measured data $\mathbf{x}$, the signal parameters $\{\mathbf{s}_i\}_{i=0}^{|\mathcal{I}|}$ corresponding to the source position and the modeled room's reflections must be estimated.

In order to make use of model (4.13) the relative distances between the microphones are required. As in our particular application the microphones are located on the loudspeakers (sources) (see Fig. 4.5), it can be assumed that the position of the source is known due to a prior calibration step [40][7]. As a result, only the positions of the first order reflections $\mathcal{S}_{1st} = \{\mathbf{s}_i\}_{i=1}^{|\mathcal{I}|}$ need to be estimated. That is, the modified data model, after removing the direct path contribution, is now given by

$$\tilde{\mathbf{x}} = \sum_{i=1}^{|\mathcal{I}|} \gamma_i \mathbf{H}(\mathbf{s}_i)\boldsymbol{\alpha} + \mathbf{w}. \tag{4.14}$$

Despite that the setup depicted in Fig. 4.5 allocates the microphones in a plane, i.e., $\mathrm{affdim}(\mathbf{E}) = 2$, and only horizontal reflections can be found, the method presented here can be applied to 3D microphones arrays which allow the localization of the reflections of the floor and ceiling.

In general, the estimation problem from (4.14) can be solved by using a nonlinear squares (NLS) estimator given by

$$\{\hat{\mathcal{S}}_{1st}, \hat{\boldsymbol{\gamma}}\} = \underset{\{\mathcal{S}_{1st},\boldsymbol{\gamma}\}}{\arg\min} \, \|\tilde{\mathbf{x}} - \sum_{i=1}^{|\mathcal{I}|} \gamma_i \mathbf{H}(\mathbf{s}_i)\boldsymbol{\alpha}\|_2^2, \tag{4.15}$$

where $\boldsymbol{\gamma} = [\gamma_1, \ldots, \gamma_{|\mathcal{I}|}]^T$ is the attenuation parameter vector. The expression in (4.15) would yield the maximum likelihood estimates [27] of the image sources locations if the noise vector $\mathbf{w}$ is both spatially and temporally white and Gaussian.

To minimize the expression in (4.15), consider first minimizing with respect $\boldsymbol{\gamma}$ for fixed $\mathcal{S}_{1st}$. This yields a new minimization where $\boldsymbol{\gamma}$ has been concentrated out. That is,

$$\hat{\mathcal{S}}_{1st} = \underset{\mathcal{S}_{1st}}{\arg\min} \, \|\Pi_{\mathbf{H}}^{\perp}\tilde{\mathbf{x}}\|_2^2, \tag{4.16}$$

where

$$\Pi_{\mathbf{H}}^{\perp} = \mathbf{I} - \mathbf{H}(\mathbf{H}^H\mathbf{H})^{-1}\mathbf{H}^H. \tag{4.17}$$

While the estimator in (4.16) provides optimal performance [54], in terms of estimation variance under white Gaussian noise, notice that it is a highly nonlinear $3|\mathcal{I}|$-dimensional optimization problem. Due to its noncovexity and high dimensionality, (4.16) becomes very expensive to implement in practice. Therefore, the following sections approach the estimation problem in (4.15) by decoupling it into a sequence of simpler optimization problems.

4.4 Wideband CLEAN

The original CLEAN algorithm was first introduced by Högbom in [22] for applications in radio astronomy. In his paper, Högbom devised CLEAN to deconvolve sky intensity distributions under the assumption that the target space can be represented by a sparse set of point sources. CLEAN considers the intensity map obtained via the delay-and-sum (DAS) beamformer [4] as the result of convolving the true source intensity with

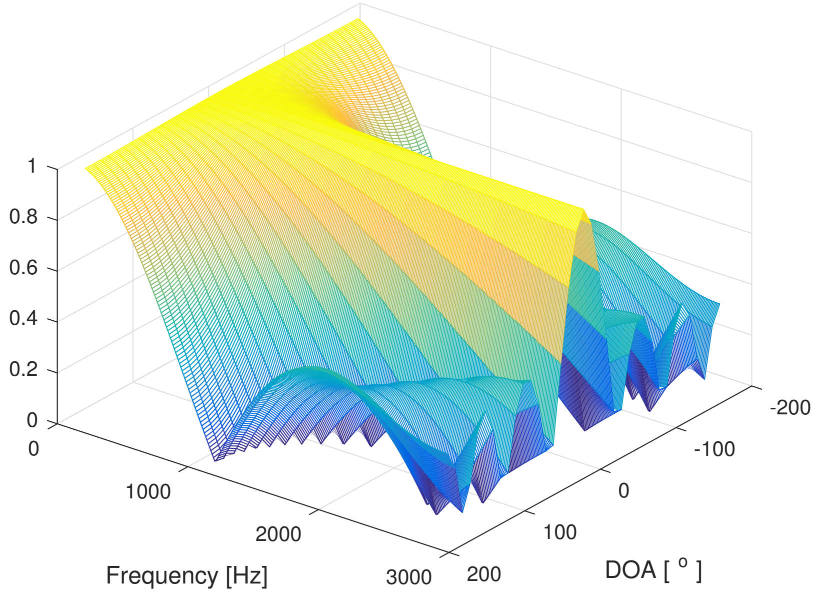


Figure 4.6: Beam pattern for a six-element UCA with 6cm of radius.

the beam pattern. Due to finiteness and irregularities in the sampling process, the beam pattern contains undesirable sidelobes. Hence, the intensity map from the DAS beamformer (often referred as "dirty map" [22]) is polluted with contributions from the sidelobes. An example of the DAS beampattern for a six-element uniform circular array (UCA) with radius of 6cm is shown in Fig. 4.6

The CLEAN algorithm aims to deconvolve the dirty map to produce an estimate of the true source intensity distribution by means of a decoupled iterative approach. In posterior work [33][46][52], extensions and modifications for the wideband signal and multiple-snapshot case have been made. In Fig. 4.7 the beampattern from the previous UCA is depicted when the data contains two sources at the directions-of-arrival (DOAs) $\boldsymbol{\theta} = [-\pi/3, \pi/3]^T$. Due to the varying mainlobe width as a function of frequency, the performance of the typical DAS beamforming method for wideband signals degrades. The mainlobe's problem is also present in adaptive algorithms [13][19] even though these methods present much better interference suppression and enhanced resolution. Moreover, the small number of segments to process and highly correlated sources, as in reverberant conditions, degrade even more the performance of adaptive algorithms. On the other hand, CLEAN eliminates the varying mainlobe width and sidelobes problems suffered from the DAS-based approaches. In addition, the single snapshot case, infeasible for adaptive algorithms, can be addressed without problems by CLEAN.

Motivated by these issues, in the following, the CLEAN algorithm described in [52] is used to solve (4.15). First, let us introduce the modified observed signal vector

$$\tilde{\mathbf{x}}_r = \tilde{\mathbf{x}} - \sum_{\substack{i=1 \\ i \neq r}}^{R} \hat{\gamma}_i \mathbf{H}(\hat{\mathbf{s}}_i) \boldsymbol{\alpha}, \tag{4.18}$$

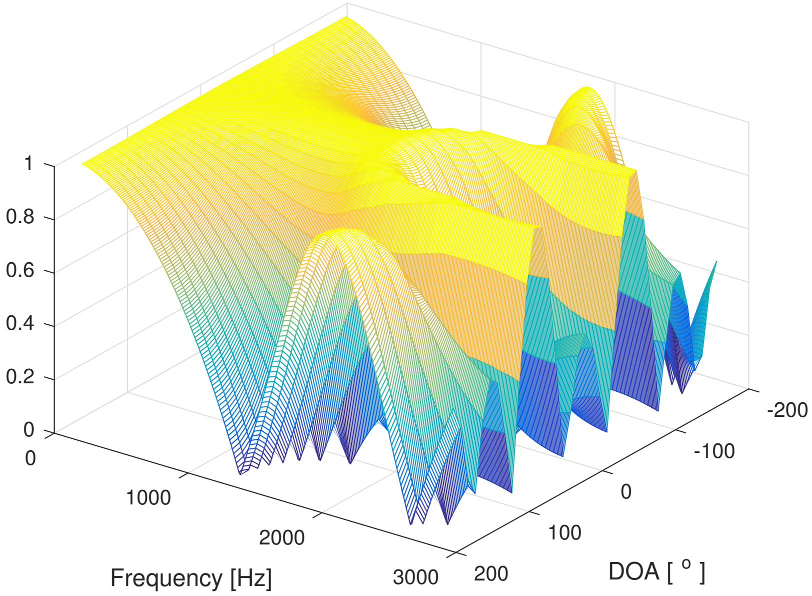


Figure 4.7: Beam pattern for a six-element UCA with 6cm of radius when two sources at $\boldsymbol{\theta} = [-\pi/3, \pi/3]^T$ are present.

where the estimates $\{\hat{\mathbf{s}}_i, \hat{\gamma}_i\}_{i=1,i\neq r}^{R}$ are assumed to be known, and $R \leq |\mathcal{I}|$. Therefore, the parameters estimate γ_r and $\mathbf{s}_r$ can be estimated using simpler estimators, i.e.,

$$\hat{\gamma}_r = \frac{\text{Re}\{\boldsymbol{\alpha}^H \mathbf{H}^\dagger(\mathbf{s}_r)\tilde{\mathbf{x}}_r\}}{\|\boldsymbol{\alpha}\|_2^2}, \tag{4.19}$$

$$\hat{\mathbf{s}}_r = \underset{\mathbf{s}_r}{\arg\min} \, \|\tilde{\mathbf{x}}_r - \hat{\gamma}_r \mathbf{H}(\mathbf{s}_r)\boldsymbol{\alpha}\|_2^2. \tag{4.20}$$

The expression from (4.19) and (4.20) estimate the positions of the $|\mathcal{I}|$ image sources disjointly, leading to a conceptually and computationally simpler solution than the one from (4.15).

The CLEAN algorithm can be summarized as follows

- **Step 1:** Initialize with r=1, $R = 1$ and let $\tilde{\mathbf{x}}_1 = \tilde{\mathbf{x}}$.
- **Step 2:** Estimate γ_i and $\mathbf{s}_r$ by using (4.19) and (4.20).
- **Step 3:** Compute the contribution of the estimated source. Subtract its contribution from $\tilde{\mathbf{x}}_r$ to produce $\tilde{\mathbf{x}}_{r+1}$ as in (4.18).
- **Step 4:** Add the current estimated signal location to the list of estimated signals and increase R by one.
- **Step 5:** Increase the iteration index r by one.
- **Remaining Steps**: Repeat Steps 2 to 5 until $r = R_{max}$ or the process works down to the noise level.

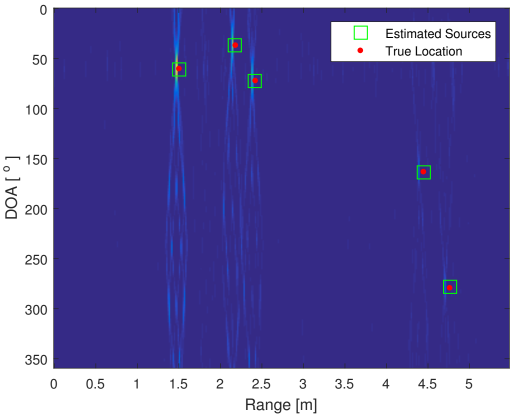


Figure 4.8: CLEAN non-linear cost function in range and DOA

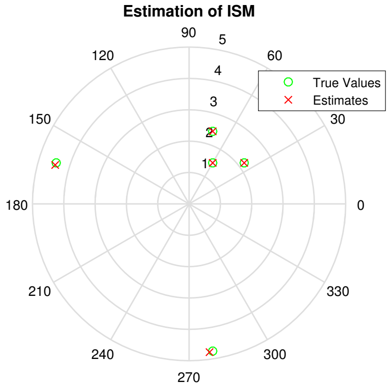


Figure 4.9: Estimated locations of the source and the image sources using CLEAN

If it is desired to apply CLEAN under the assumption of point sources, R_{max} should represent the maximum number of reflections of interest, i.e., $R_{max} = |\mathcal{I}|$. Otherwise, CLEAN works down the dirty map until it reaches the noise floor. This fact makes the CLEAN algorithm suitable for imaging distributed sources [52]. An example for source localization using CLEAN is shown in Fig. 4.8 and Fig. 4.9. The left figure shows the cost function for a single six-microphone UCA, with white noise as transmitted signal and a sampling frequency of 20kHz.

Finally, note that the feasible set of the optimization problem in (4.20), for the flat wall case, can be reduced by making the following considerations (see Fig. 4.10):

1. As the position of the true source $\mathbf{s}$ is known, the feasible set excludes all the points belonging to the set

$$\mathcal{A} = \bigcup_{m=1}^{M} \mathcal{B}_m \tag{4.21}$$

   where $\mathcal{B}_m = \{\mathbf{p} : \|\mathbf{r}_m - \mathbf{p}\| < \|\mathbf{s} - \mathbf{r}_m\|_2\}$.

2. In consecutive iterations the feasible set excludes the points *behind*, i.e. positive inner product, with respect to the boundary plane constructed using (2.14) defining a room wall.

4.5 Wideband RELAX

An alternative to the CLEAN algorithm, is the RELAX method. Instead of working down the dirty map until the noise floor or the maximum number of sources in a sequential fashion, RELAX uses the knowledge of estimated sources to improve previous estimates.

Originally, RELAX was first proposed in [33] as a asymptotically efficient method for mixed spectrum estimation. Later, RELAX was extended for other applications such as DOA and waveform estimation [32]. Finally, in [52] an extension for wideband

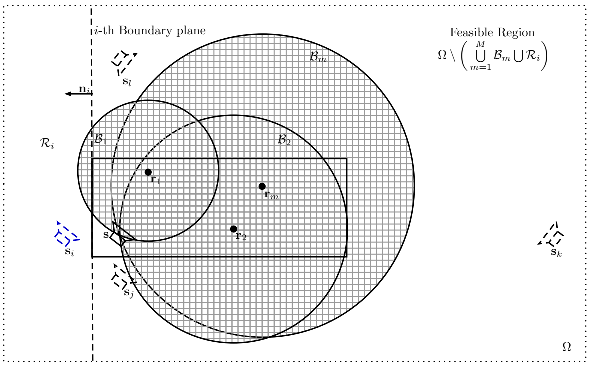


Figure 4.10: Illustration of the reduction in the feasible set for the image source localization problem. The already estimated source $\mathbf{s}_i$ (blue) defines the first boundary of the room.

signals was applied in aeroacoustic imaging. In general terms, the RELAX method can be seen as an improved version of CLEAN, which trade computational complexity for estimation performance. The difference between these two methods lays in an inner loop that RELAX adds in order to refine the first estimates. Similar to CLEAN, RELAX does not suffer from effects of varying mainlobe and sidelobe problems[33].

Using the estimators from (4.19) and (4.20) and the update equation (4.18), the RELAX algorithm can be summarized as follows

- **Step 1:** Assume $R = 1$. Estimate $\{\hat{\gamma}_r, \hat{\mathbf{s}}_r\}_{r=1}$ using (4.19) and (4.20) with $\tilde{\mathbf{x}}_1 = \tilde{\mathbf{x}}$.
- **Step 2:** Assume $R = 2$. Compute $\tilde{\mathbf{x}}_2$ with (4.18) using $\{\hat{\gamma}_r, \hat{\mathbf{p}}_r\}_{r=1}$ obtained in the previous step. Estimate $\{\hat{\gamma}_r, \hat{\mathbf{s}}_r\}_{r=2}$ from $\tilde{\mathbf{x}}_2$. Next, compute $\tilde{\mathbf{x}}_1$ using (4.18) with $\{\hat{\gamma}_r, \hat{\mathbf{s}}_r\}_{r=2}$, and redetermine $\{\hat{\gamma}_r, \hat{\mathbf{s}}_r\}_{r=1}$ using (4.19) and (4.20). Repeat the step until convergence of the estimates $\{\hat{\gamma}_r, \hat{\mathbf{s}}_r\}_{r=1,2}$ or the maximum number of iterations is reached.
- **Step 3**: Assume $R = 3$. Compute $\tilde{\mathbf{x}}_3$ using the estimates $\{\hat{\gamma}_r, \hat{\mathbf{s}}_r\}_{r=1,2}$ from the previous step. Estimate $\{\hat{\gamma}_r, \hat{\mathbf{s}}_r\}_{r=3}$ from $\tilde{\mathbf{x}}_3$. Then, re-estimate $\{\hat{\gamma}_r, \hat{\mathbf{s}}_r\}_{r=1}$ from $\tilde{\mathbf{x}}_1$ computed using $\{\hat{\gamma}_r, \hat{\mathbf{s}}_r\}_{r=2,3}$. Repeat these substeps until convergence of estimates or maximum number of iterations is reached.
- **Remaining Steps**: Continue similarly until R is equal to maximum number of sources to estimate.

The convergence of the estimates can be checked by computing the cost function at the i-th iteration of each step as

$$J(i) = \|\tilde{\mathbf{x}} - \sum_{r=1}^{R} \gamma_r \mathbf{H}(\mathbf{s}_r) \boldsymbol{\alpha}\|_2^2. \tag{4.22}$$

When the change in the cost function between two consecutive iterations is smaller than a given threshold, i.e., $\|J(i) - J(i-1)\|_2^2 \leq \epsilon$, practical convergence is reached. Similar to CLEAN, the feasible set for RELAX can be reduced by applying the considerations discussed in the previous section.

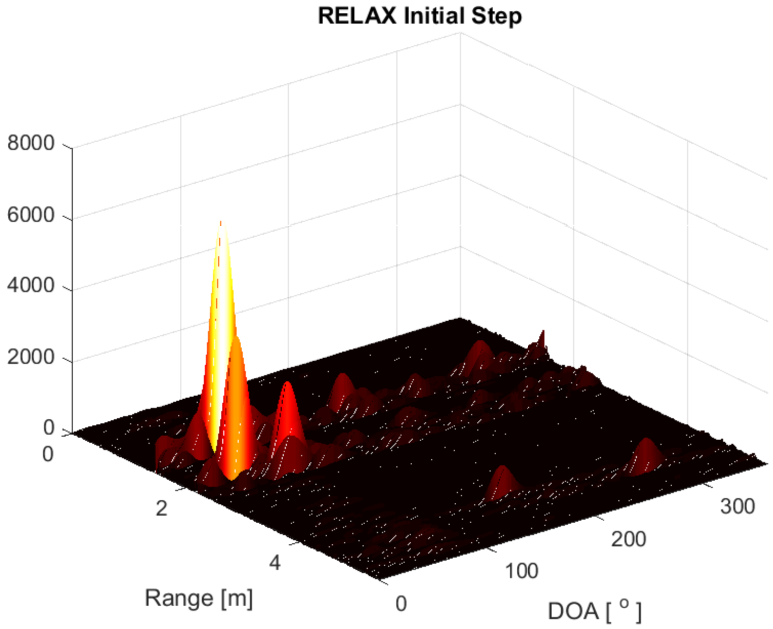


Figure 4.11: Cost function at Step 1 of RELAX using a six-element UCA with radius of 6cm. The simulated scene contained a source and four wall reflections.

An example of the output for the iterative processes described is shown in Fig. 4.11 and Fig. 4.12. The source and image sources locations are estimated using the RELAX algorithm and six-microphone UCA with radius of 6cm. Only the four reflections are considered in the simulation, i.e., ceiling and floor were considered covered with absorbent material. From Fig. 4.12 can be noticed how RELAX identify some of the weakest reflections that were shadowed by the sidelobes of stronger reflections. Fig. 4.13 and Fig. 4.14 illustrates the difference in the estimation between CLEAN and RELAX. In this case, CLEAN exhibits a heavy bias in one of the image sources, leading to a poor estimation of its position. However, due to its multiple iterations, RELAX is able to properly find the source and the four reflections in the scene. The simulation was made considering three 3-microphones UCAs with a radius of 6cm randomly placed in a shoe-box shaped room. The sampling frequency used was fs = 96kHz. A measured loudspeaker transfer function is used to include the effects of the employed driver. The response of the loudspeaker is considered to be omnidirectional.

4.6 Theoretical performance

It is of interest to understand how accurate the image sources locations can be estimated. In order to do so, the Cramér-Rao lower bound (CRLB) [27] is used to provide a theoretical lower bound for the variance of any unbiased estimator. Appendix D derives the Fisher information closed expression for the cases of single source and multiple sources in white Gaussian noise under the assumption of near and far-field propagation and isotropic radiation.

From Appendix D, the (i, j)-th entry for the Fisher information in the case of a

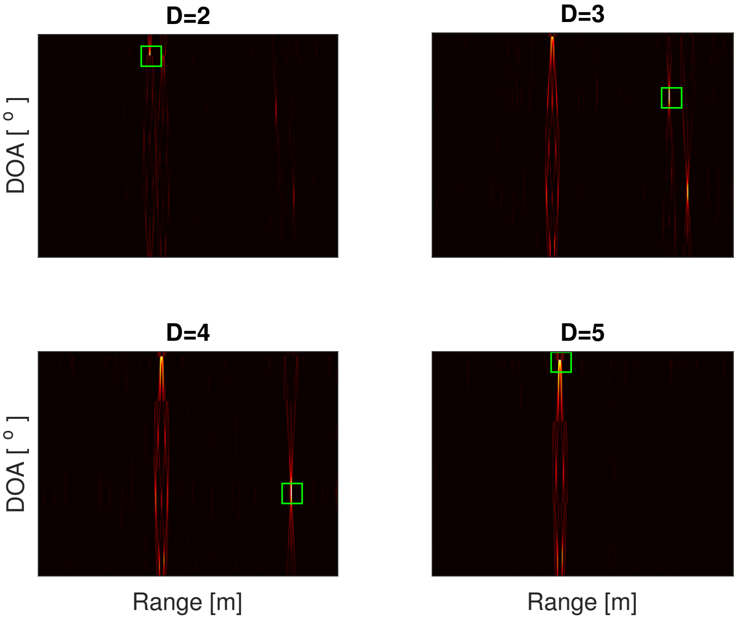


Figure 4.12: Iterative process for RELAX. In each step the number of sources increases and recalculation of previous estimates occurs.

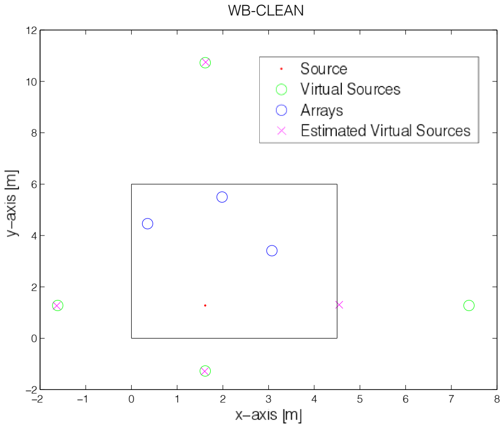


Figure 4.13: CLEAN estimates

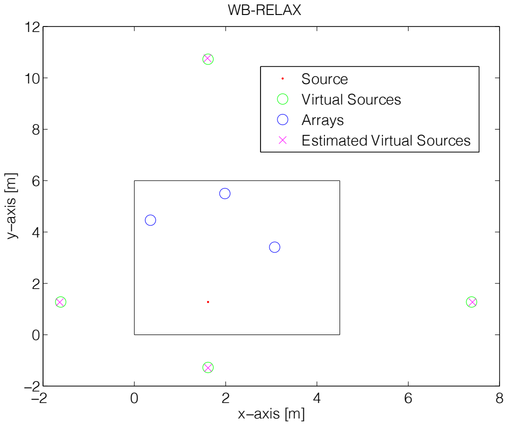


Figure 4.14: RELAX estimates

single source in the near field of an array is given by (D.14)

$$\left[J(\mathbf{s})\right]_{ij} = \frac{T}{\pi} SNR \Bigg[B \sum_{m=1}^{M} \frac{[\mathbf{r}_m - \mathbf{s}]_i [\mathbf{r}_m - \mathbf{s}]_j}{\|\mathbf{s} - \mathbf{r}_m\|^6} + (B\omega_c^2 + B^3/12) c^{-2} \sum_{m=1}^{M} \frac{[\mathbf{r}_m - \mathbf{s}]_i [\mathbf{r}_m - \mathbf{s}]_j}{\|\mathbf{s} - \mathbf{r}_m\|^4} \Bigg], \quad (4.23)$$

where $J(\mathbf{s}) \in \mathbb{R}^{D \times D}$ is the Fisher information matrix and $\mathbf{s} \in \mathbb{R}^D$ the parameter vector containing the spatial location of the source. This Fisher information has been derived

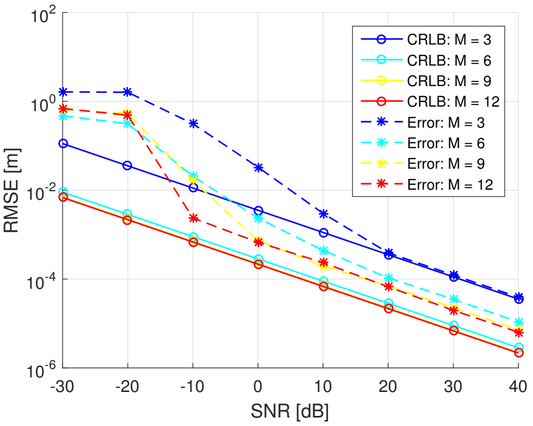


Figure 4.15: Influence of M in estimation accuracy

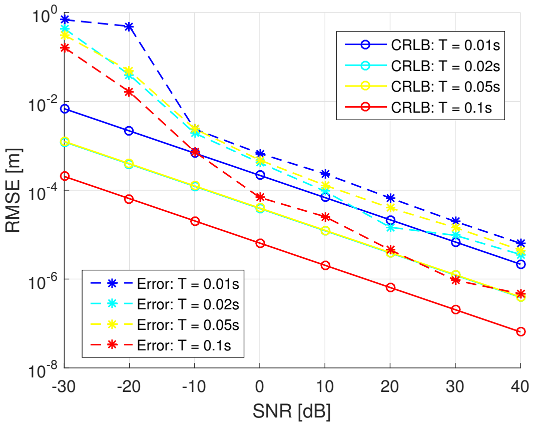


Figure 4.16: Influence of T in estimation accuracy

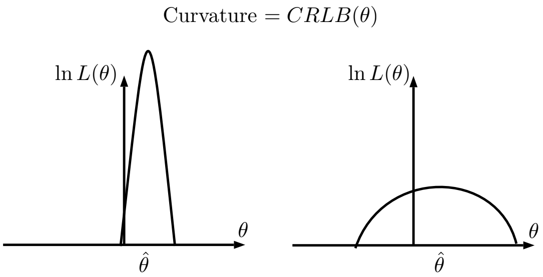


Figure 4.17: Illustration of curvature of likelihood function $L(\theta)$ for cases with large Fisher information (left) and low Fisher information (right)

considering a signal with constant power inside the frequency band $[\omega_c - B/2, \omega_c + B/2]$rads and with a duration of T seconds.

The expression in (4.23) provides great insight into the parameters that play a role in the estimation accuracy. As the CRLB is given by [27]

$$CRLB(\mathbf{s}) = J(\mathbf{s})^{-1}, \tag{4.24}$$

it is seen that by increasing the following parameters, besides the SNR, the $CRLB(\mathbf{s})$ can be reduced

- Listening time (T)
- Bandwidth (B)
- Central frequency (ω_c)
- Number of microphones (M)

To illustrate the influence of these parameters, and some of the issues that might arise using a MLE, experiments are carried out for the problem of estimating the source position when additive noise is present. In Fig. 4.15 and Fig. 4.16 the effect of number of microphones (M) and listening time (T) is shown respectively. By increasing

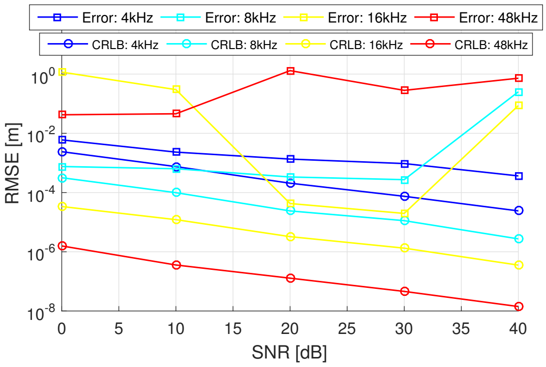


Figure 4.18: Influence of B in estimation accuracy for fix grid size

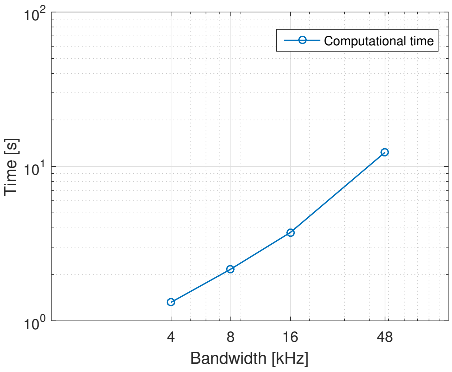


Figure 4.19: Computational complexity as function of B

the number of microphones and listening time the CRLBs decrease. In addition to the CRLBs, the root mean square error (RMSE) of the estimator from (4.15) is shown. The non-uniform spacing between the CRLBs lines is due to the weighting of the random located microphones, i.e., the sums in (4.23) have in the denominator the distance between microphone and source. These simulations were performed using a uniform grid with spacing $h = 5$cm for obtaining a coarse estimate. Posterior refinement of the estimate is done by using a derivative-free method [30].

A clear trade-off for fixed $T = N/f_s$ occurs when the sampling frequency of the system has to be selected. If all parameters are held fixed, in order to provide the same performance guarantees when the sampling frequency f_s increases it is required that the number of samples N also increases. These increments the computational burden of the estimation method.

As signals with constant magnitude across a frequency band are being considered, the effect of B and ω_c can be observed jointly. In Fig. 4.18 the RMSE of the estimation is shown for different bandwidths. The central frequency is selected as $\omega_c = B/2$. By observing the CRLBs it is expected that an increase in B leads to a decrease of the RMSE. However, Fig. 4.18 shows that this is not necessary the case when the same estimation grid is used for all bandwidths. Higher bandwidths show worse performance than lower bandwidths when the grid size is fixed to $h = 5$cm. This behavior can be explained by using intuition derived from the CRLB itself (see Fig. 4.17). As B increases, the curvature of the likelihood function, defining the local spread around the parameter of interest, becomes much more pronounced. As a result, in order to avoid getting caught in a local minimizer of (4.15) a finer grid is required. Strategies to adjust the grid spacing in different scenarios have been discussed in [14] and [39]. Taking this issue in consideration, in Fig. 4.20 the RMSE is shown for increasing bandwidths with non-fix grid size. When an appropriate grid size is used for the estimation task, it is seen how the decrease in error follows the trend of the CRLB. However, this improvement in performance increases the computational cost (orange curve) of the method. The grids size used are [5e−2, 2e−2, 1e−2, 3e−3]m for the bandwidths [4, 8, 16, 48]kHz respectively.

Similar to the single source case, the multiple source scenario, e.g., set of first-order reflections, is examined using the CRLB as theoretical limit for our estimators.

From the results in Appendix D, the (k, l)-th entry of the (i, j)-th block of the Fisher information $J(\mathcal{S}_{1st}) \in \mathbb{R}^{|\mathcal{I}|D \times |\mathcal{I}|D}$ is given by (D.34)

$$\left[J_{ij}(\mathcal{S}_{1st})\right]_{kl} = \frac{T}{\pi} SNR \sum_{m=1}^{M} A_{ij}^{(m)}(k,l) \Bigg[B_{ij}^{(m)} \int\limits_{\omega_c - B/2}^{\omega_c + B/2} \cos(\omega \alpha_{ij}^{(m)}) d\omega + c^{-2} \int\limits_{\omega_c - B/2}^{\omega_c + B/2} \omega^2 \cos(\omega \alpha_{ij}^{(m)} d\omega) - c^{-1} C_{ij}^{(m)} \int\limits_{\omega_c - B/2}^{\omega_c + B/2} \omega \sin(\omega \alpha_{ij}^{(m)}) d\omega \Bigg], \quad (4.25)$$

where $\alpha_{ij}^{(m)}$, $A_{ij}^{(m)}(k,l)$, $B_{ij}^{(m)}$ and $C_{ij}^{(m)}$ have been defined in (D.28)-(D.31) and are functions of the distances between the m-th microphone and the i-th and j-th sources.

Observing (4.25), it is noticed that the parameters that influence the Fisher information in the single source case also affect the Fisher information in the multiple source instance. This is an expected behavior as the diagonals blocks, i.e., $i = j$, of the Fisher information matrix are the Fisher information matrix of the single source case (see Appendix D).

To illustrate the difference between CLEAN and RELAX a set of Monte Carlo simulations is performed. The intention of these experiments is to observe whether the estimators, for the multiple sources case, are capable of attaining the CRLB. The simulation setup is shown in Fig. 4.21. The scene depicts the situation of a set of loudspeakers equipped with three-element UCAs of 6cm radius. In this setup, the goal is to estimate the positions of the first-order reflections created by one of the loudspeakers using the other three, i.e., $M = 9$ available microphones. The loudspeaker is considered an isotropic source emitting a signal of length $N = 801$ samples with flat spectrum of bandwidth $B/2\pi = 4$kHz and central frequency $\omega_c = B/2$.

From Fig. 4.22 it is seen that both methods present the same performance for low SNR, however after the region of low SNR is crossed, CLEAN is not able to follow the CRLB trend while RELAX continues decreasing its RMSE. In addition, Fig. 4.23 shows a comparison between the computational cost of the methods for the different SNR.

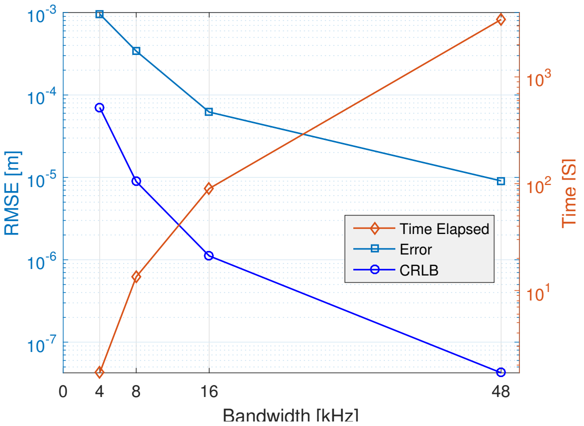


Figure 4.20: Influence of B in estimation accuracy for adjusted grid sizes

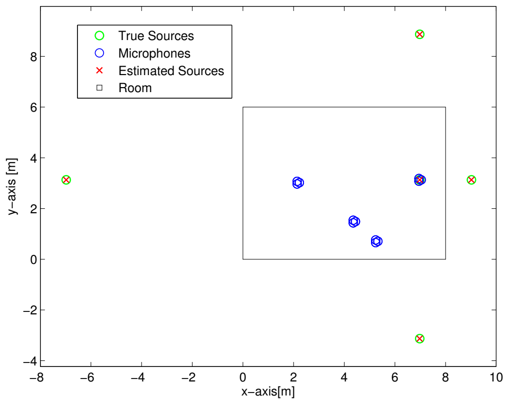


Figure 4.21: Simulated scene for estimation of first-order reflections

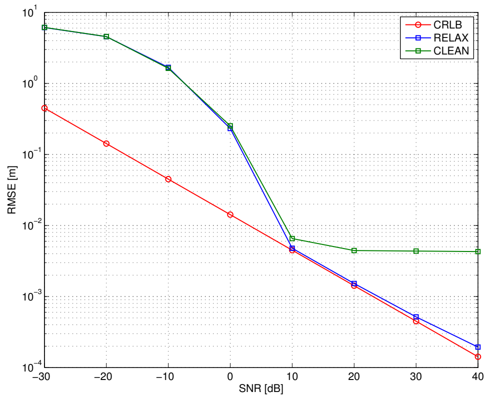


Figure 4.22: Performance comparison between CLEAN and RELAX

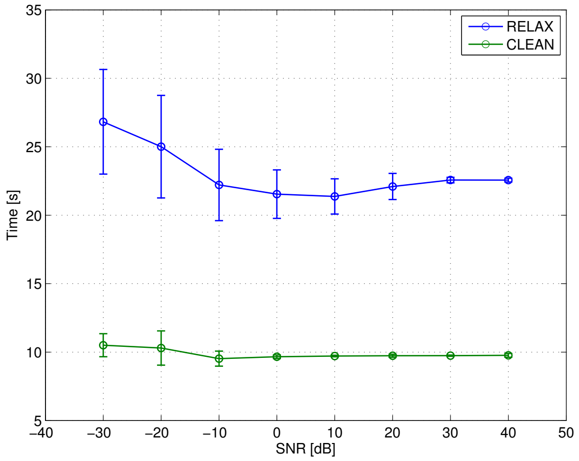


Figure 4.23: Comparison of computational load of CLEAN and RELAX

The reduced bias of RELAX requires at least twice the computational time of CLEAN. Notice that while CLEAN remains stable for all the SNR range, RELAX performs more iterations at low SNR. However, these iterations do not provide any improvement as the results in Fig. 4.22 indicate that CLEAN and RELAX obtain a similar performance in this region.

To provide an explanation to the difference in the high SNR regime between CLEAN and RELAX, let us look closer into the mean square error (MSE) of our estimators. Following the MSE definition for an estimate $\hat{\theta}$ [27]

$$MSE(\hat{\theta}) = E\big[(\hat{\theta} - \theta)^2\big], \tag{4.26}$$

it is possible to express the MSE in terms of its components as

$$\begin{aligned} MSE(\hat{\theta}) &= E\Big[\big(\hat{\theta} - E(\hat{\theta})\big)^2\Big] + \big(E(\hat{\theta}) - \theta\big)^2 && (4.27)\\ &= \mathrm{Var}(\hat{\theta}) + \mathrm{Bias}(\hat{\theta}, \theta)^2. && (4.28) \end{aligned}$$

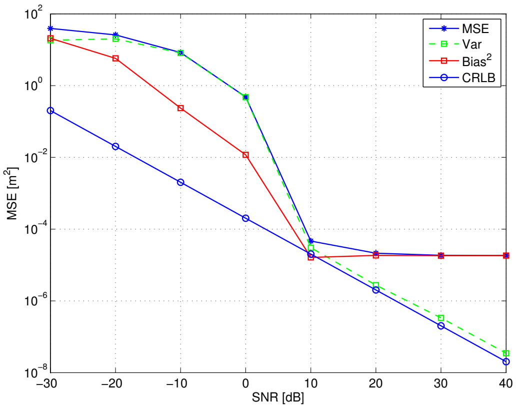


Figure 4.24: Variance-Bias of CLEAN estimator

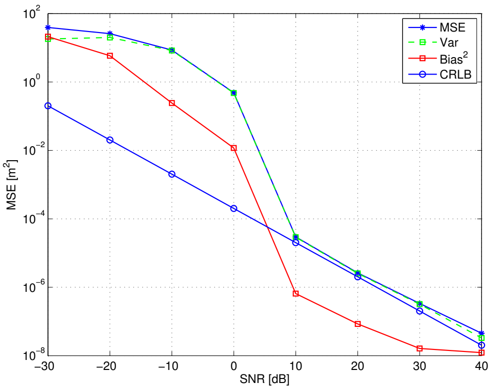


Figure 4.25: Variance-Bias of RELAX estimator

The CRLB is a lower bound for the variance of an unbiased estimator, hence if the methods are indeed unbiased estimators of the unknown parameter vector, their MSEs cannot be lower than the CRLB. In Fig. 4.24 and Fig. 4.25 both terms in (4.28) are computed for each estimator. It can be observed that CLEAN is heavily biased compared to RELAX. Hence, even though CLEAN variance continuously decreases, its MSE gets stuck due to its built-in bias. The bias in CLEAN is a well-known result in radioastronomy literature [49][35], where CLEAN has been recognized as a matching pursuit algorithm [34]. The bias in RELAX (Fig. 4.25), which becomes evident at higher SNR is considered to be due to the tolerance of the derivative-free optimization method used to refine the coarse estimates.

4.6.1 On the selection of parameters

To finalize the theoretical evaluation of the methods, the selection of the parameters affecting the CRLB of the estimates is discussed.

The model used to describe the propagation of the emitted acoustic signal considers isotropic radiators. However, in practice, a loudspeaker can not be considered as one for the whole frequency range. This is because the directivity pattern of a loudspeaker depends on the relation between the size of the loudspeaker and the wavelength of the reproduced sound. Therefore, it is expected that the directivity of the loudspeaker degrades the performance of the methods discussed in this chapter. This problem is illustrated in Fig. 4.26 where the effect of the beam pattern in the intensity of the image sources is shown. Assuming a loudspeaker with irregular directivity pattern for a single frequency, as depicted in Fig. 4.26, it can be seen that the image sources, representing each reflection from the walls, are attenuated differently. In this particular case, as the loudspeaker has almost no radiation in its rear direction, the image source behind the loudspeaker is severely attenuated. This attenuation is combined with the attenuation due to path length and wall absorption reducing the SNR of our available data. As a result, the estimation performance of the image sources degrades.

To alleviate this problem, the loudspeaker can be used in a region where it has low directivity. This not only provides a better distribution of the radiation pattern, but

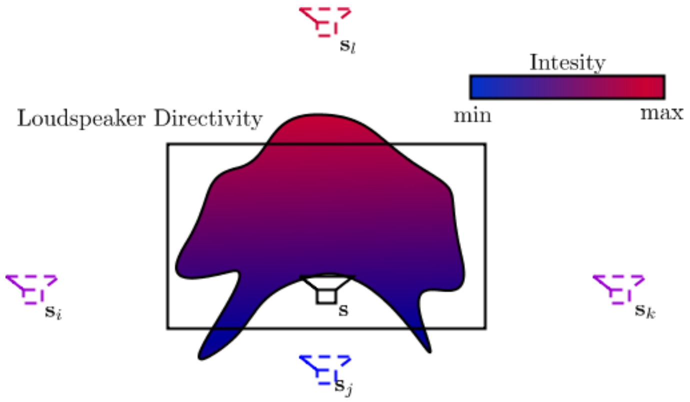


Figure 4.26: Effect of loudspeaker directivity in image sources

also allows us to have a coarser estimation grid which results in lower computational complexity. Furthermore, due to the low frequency region of operation, a low f_s suffices to sample the process. This reduces the number of samples N needed to be able to receive the contribution of all image sources, i.e.,

$$N \geq \left\lceil \frac{d_{\max} f_s}{c} \right\rceil, \tag{4.29}$$

where c is the speed of sound and $d_{\max}$ is the largest distance between any microphone and any image source. Expression (4.29) implies that the number of samples depends on the size of the room which boundaries are going to be estimated. In this work, as a good compromise between loudspeaker directivity, computational complexity, and validity of the geometrical acoustic model, frequencies $\mathcal{F} \leq 4$kHz are recommended for performing image source estimation.

In the case of the number of microphones, following the CRLB, more does not necessary imply better. Even though increasing the number of microphones reduces the CRLB, their location has a much higher impact in the estimation performance. Unfortunately, in a typical reproduction environment, where the microphones are built into on the loudspeakers, the distribution of the devices is decided by the users following aesthetic reasons. As a result, almost nothing can be done in this respect.

Although in practice the number of microphones is constrained by the number of devices distributed in the room, a typical reproduction setup with four loudspeakers provides $M = 9$ microphones for processing if a three-element UCA is included in each loudspeaker and only three of them are used to estimate the sources created by the fourth loudspeaker. From Fig. 4.15 and Fig. 4.22 it is seen that, in the ideal case, accuracy below centimeters can be achieved using this number of receivers.

Finally, when in principle these methods can use an arbitrary signal to estimate the image source locations, it is recommended to employ zero mean white Gaussian noise as probing signal $s(t)$. The reason behind this is that its autocorrelation function is the delta function, i.e.,

$$r(\tau) = E\big[s(t)s(t-\tau)\big] = \delta(\tau), \tag{4.30}$$

which provides the best performance for identifying delays between shifted signals [28].

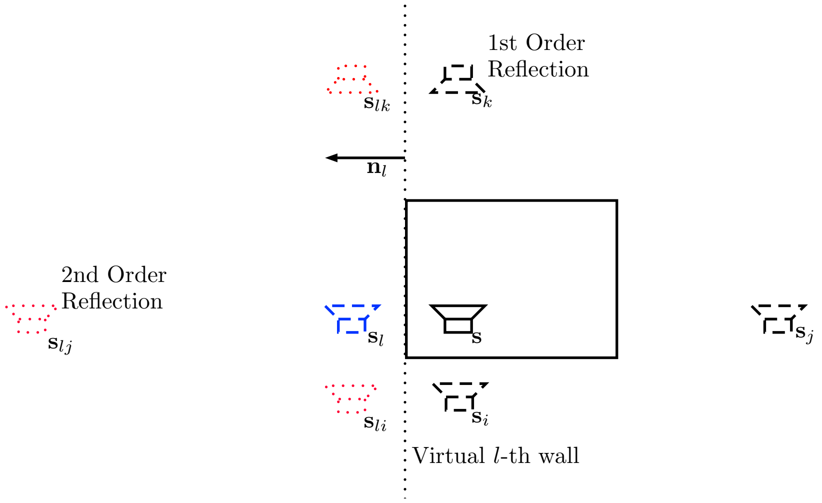


Figure 4.27: Second order reflections in room with respect the l-th wall

4.7 Inclusion of higher order reflections

Given that from a set of reflective walls it is possible to compute any reflection of a given order in a shoe-box shaped room, an alternative estimation procedure for the estimation of the first order reflections is proposed in this section.

Assuming the image source model for modeling the room's reflections and a shoe-box shaped room, every time a new reflection is estimated knowledge of higher-order reflections can be included in the minimization problem (4.15) to refine the estimate of the reflection location. To illustrate this, consider the 2D case shown in Fig. 4.27. After $\mathbf{s}_l$ (blue) has been estimated, the boundary defining the l-wall (dotted) can be estimated and used to generate the 2nd-order reflections from the other image sources. Suppose the source $\mathbf{s}_r, r \in \{i, j, k\}$ is going to be estimated next. The modified observed signal vector from (4.18) is given by

$$\tilde{\mathbf{x}}_r = \gamma_r \mathbf{H}(\mathbf{s}_r)\boldsymbol{\alpha} + \sum_{\substack{n \leq R \\ n \neq r}} \gamma_{rn} \mathbf{H}(\mathbf{s}_{rn})\boldsymbol{\alpha} + \mathbf{w}_r, \tag{4.31}$$

where $\mathbf{w}_r$ contains the modeling and additive errors by the previous estimates, and the uncorrelated measurement noise. The 2nd-order reflection position $\mathbf{s}_{rn}$ can be computed by

$$\mathbf{s}_{rn} = \mathbf{s}_r + (\mathbf{s}_n - \mathbf{s}). \tag{4.32}$$

Noticing that γ_{rn} is the product of the attenuation of the r-th and n-th walls [1], (4.31) can be rewritten as

$$\tilde{\mathbf{x}}_r = \gamma_r \left(\mathbf{H}(\mathbf{s}_r) + \sum_{\substack{n \leq R \\ n \neq r}} \gamma_n \mathbf{H}(\mathbf{s}_{rn}) \right) \boldsymbol{\alpha} + \mathbf{w}_r. \tag{4.33}$$

At this point the estimates $\{\hat{\gamma}_n\}$ $\forall$ $n \leq R, n \neq r$ are available, hence to provide an estimate of $\{\gamma_r, \mathbf{s}_r\}$ the unknown attenuation coefficients can be substituted by their

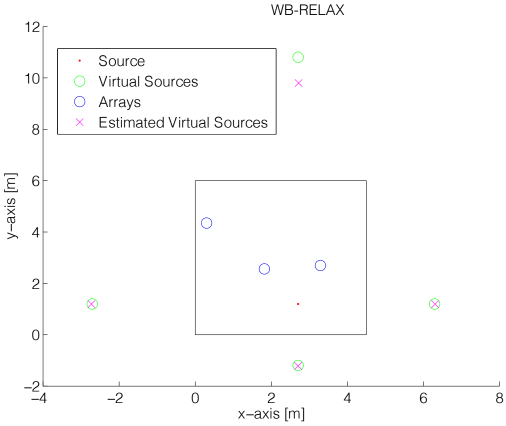


Figure 4.28: RELAX estimates

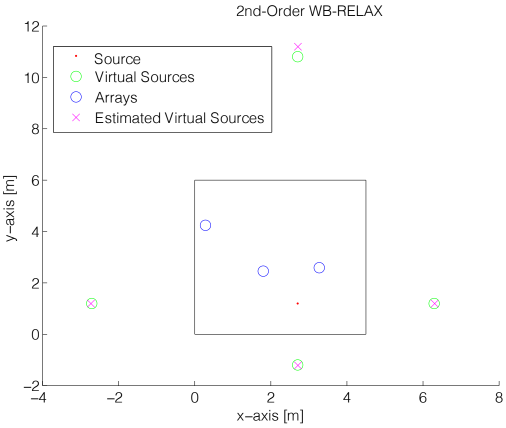


Figure 4.29: 2nd Order RELAX estimates

estimates. Therefore, the estimates for the attenuation and position for the r-th image source are given by

$$\hat{\gamma}_r = \frac{\mathrm{Re}\{\boldsymbol{\alpha}^H \bar{\mathbf{H}}^\dagger(\mathbf{s}_r)\tilde{\mathbf{x}}_r\}}{\|\boldsymbol{\alpha}\|_2^2}, \tag{4.34}$$

$$\hat{\mathbf{s}}_r = \arg\min_{\mathbf{s}_r} \|\tilde{\mathbf{x}}_r - \hat{\gamma}_r \bar{\mathbf{H}}(\mathbf{s}_r)\boldsymbol{\alpha}\|_2^2, \tag{4.35}$$

where

$$\bar{\mathbf{H}}(\mathbf{s}_r) = \mathbf{H}(\mathbf{s}_r) + \sum_{\substack{n \leq R \\ n \neq r}} \hat{\gamma}_n \mathbf{H}(\mathbf{s}_{rn}). \tag{4.36}$$

By changing the estimators in (4.19) and (4.20) for the ones in (4.34) and (4.35), the steps described in the previous sections for both CLEAN and RELAX can be followed to obtain the locations estimates for images sources. If the image model holds for the data, it is expected that the estimators taking into consideration the 2nd-order reflections achieve a better performance as in every iteration, for each step, a joint estimation process is performed, i.e., even though the contributions of the sources $\{\hat{\mathbf{s}}_n\}\ \forall\ n \leq R, n \neq r$ are removed from the data, their locations are used to estimate $\hat{\mathbf{s}}_r$.

In Fig. 4.28 and Fig. 4.29 an example comparing RELAX and RELAX using second-order reflections is shown. In this example, a white Gaussian noise signal was convolved with the measured transfer function of a loudspeaker and sampled at 96kHz. Three 3-microphone UCAs were randomly placed in the room. From Fig. 4.29 it is clear that the farthest image source is better localized when the higher-order reflections are used.

Finally, when all the image sources are known, similarly as the in echo sorting problem, the boundaries of the room can be estimated by the geometrical method described in Chapter 2.

4.8 Experimental Results

In this section results from experiments are presented to evaluate the performance of the proposed iterative methods in this chapter. Similar to the previous chapter, in this section the performance of the proposed methods is studied for varying SNR, reverberation time, and number of sources. In addition, a comparison between the EDM-based methods (discussed in the previous chapter) and the methods introduced in this chapter is presented.

A set of 500 Monte Carlo simulations are performed for each parameter subject to study placing loudspeakers equipped with microphones in a room close to walls and/or corners. This kind of configuration is used for the experiments as it is one of the most common distribution for the loudspeakers in typical audio reproduction setups. Using the position of the loudspeakers and microphones as input, for each receiver-source pair a 3D room impulse response is generated using [20]. The data is generated by convolving the simulated RIRs with a white Gaussian noise signal of 4kHz bandwidth and a measured loudspeaker impulse response. The listening time is considered as three times the largest distance between a microphone and an image source. The experiments are run in Matlab on a Macbook Air (Mid 2013) 1.7 GHz Inter Core i7 processor.

The room used in these experiments has a constant volume of $280m^3$ with dimensions $8m \times 6m \times 5m$. As the tested setup allocates the microphones in a plane, the RMSE used to quantify the methods performance is the expectation of the square root of the error squared between estimated and true $2D$ room vertices. That is,

$$\text{RMSE}(\hat{\boldsymbol{\theta}}) \triangleq \sqrt{E\left[\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2\right]}, \tag{4.37}$$

where $\boldsymbol{\theta} \in \mathbb{R}^4$ represents the $2D$ room vertices and $\hat{\boldsymbol{\theta}} \in \mathbb{R}^4$ the estimates of the vertices given by the method.

4.8.1 Effect of Loudspeaker Transfer function and Simulated RIR

So far only data following the proposed model in (4.14) has been considered. By doing so, it has been shown that the proposed RELAX estimator can be an efficient and consistent estimator, attaining the CRLB. However, the synthetic data generated following the model (4.14) represents the ideal case of our problem. That is, it does not consider that the attenuation coefficients are frequency dependent, the noise is not only white Gaussian noise, and that phase changes in the signals are induced by their reflections in the walls. Furthermore, the assumption of an ideal isotropic radiator is not valid anymore when a loudspeaker is used to reproduce sound.

To demonstrate the effect of deviations from the ideal case, simulations using the RIR generator [20] and the measured transfer function of a loudspeaker shown in Fig. 4.1 are performed. In Fig. 4.30 and Fig. 4.31 the RMSE of the estimated vertices positions vector $\hat{\boldsymbol{\theta}}$ and the computation time for the different methods are shown. For this simulation, only one loudspeaker is considered as source, and the other three are used to estimate the first-order reflections. As it would be expected, the increase of the model mismatch by using simulated 3D RIRs and coloring the signal with the loudspeaker transfer function, degrades the estimators performance. Now, instead of sub-

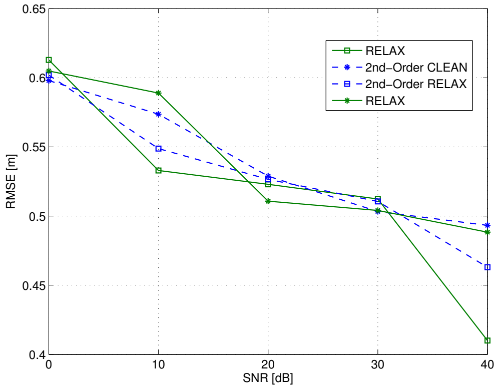


Figure 4.30: Estimation results from the different methods when the loudspeaker transfer function is considered

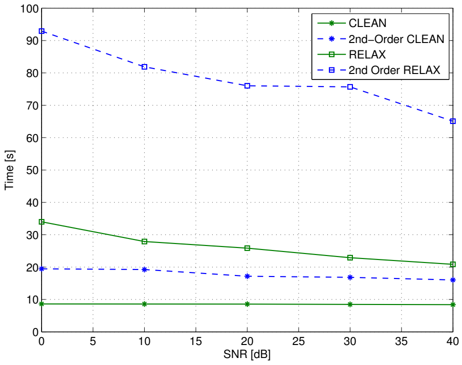


Figure 4.31: Computation time of the different methods

centimeter accuracy as in Fig. 4.22, due to the model mismatch with respect to (4.14) our estimation error is now around half a meter. In addition, from Fig. 4.30 is seen that the four methods performance very similarly. However, the computation time for the RELAX-based methods is higher than for the CLEAN-methods. Furthermore, a large gap between in the computational time 2nd-Order RELAX and RELAX is found. By observing the estimation results, it is suggested that the additional computational cost from evaluating the modified cost function does not provide in average significant improvements in terms of estimation accuracy.

4.8.2 Reverberation Time of the Room

Reverberation can be generally defined as the persistence of sound after a sound is produced, and it is present in almost all situations where the first-order reflections are to be estimated. Furthermore, it is well-known that it causes negative effects in source localization methods, as it contributes with correlated noise. These effects lead to degradation on the performance of the proposed methods.

Considering the same scenario, i.e., one loudspeaker as source and three loudspeakers used for estimate the first-order reflections, a set of simulations with a SNR of 40dB are performed for different reverberation times ($T60$). The RMSE of the estimated vertices positions vector and its variance are shown in Fig. 4.32 and Fig. 4.33. In general, an increasing trend in the estimation error can be observed from these results, particularly the variance of our estimates seems to steadily increase with higher reverberation time. As reverberation adds correlated noise, a worse performance is expected with respect the uncorrelated noise case. This correlated noise not only spreads our estimation, as white Gaussian noise, but in addition it adds bias to our estimate depending on the particular realization of the reverberation tail, i.e. as the RIR simulator generates the reverberation tail of the RIR using a stochastic method, all RIR realizations are inherently different.

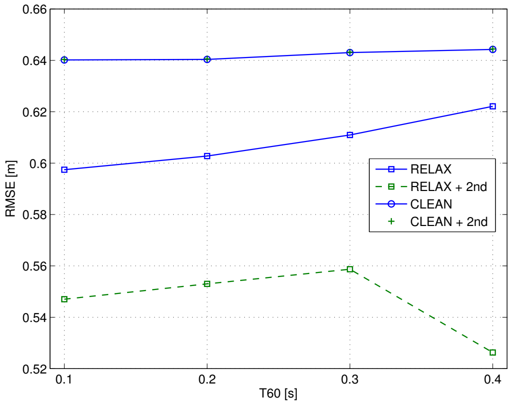


Figure 4.32: Room reconstruction error comparison for different reveberation times

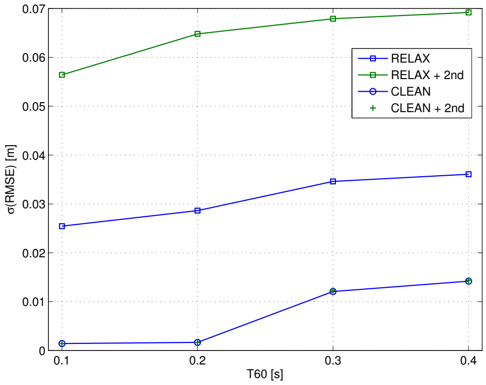


Figure 4.33: Standard deviation of reconstruction error for different reverberation times

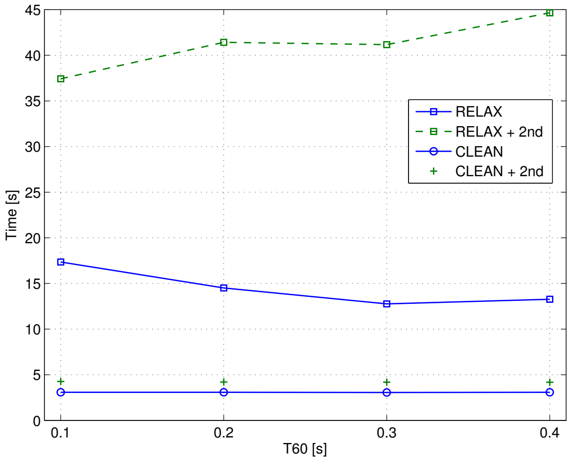


Figure 4.34: Computation time as function of reverberation time

4.8.3 Number of Sources

As in this setup more than one loudspeaker is available to produce measurements, the effect of the number of sources on the estimation error is evaluated. In this experiment, a shoe-box shaped room with $T60 = 0.2$ and a SNR of 40dB is considered. In Fig. 4.35 and Fig. 4.36 the reconstruction error and the computation time for each of the methods are shown respectively.

From each source a set of vertices locations are estimated from the first-order reflections. The vertices are then grouped on clusters and the centroid of each cluster is considered as the final vertices locations estimate. Adding more sources decreases the estimation error as we compensate for estimation errors of the first-order reflections. However, the computation time increases almost linearly, being the RELAX-based methods the ones with the highest computational time. In general, adding more sources further decreases our estimation error, however this is constrained to the num-

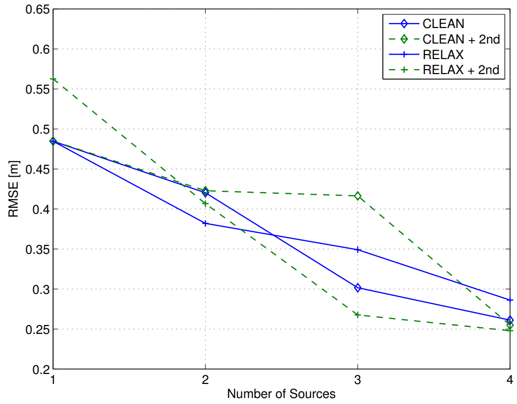


Figure 4.35: Room reconstruction error comparison for different number of sources

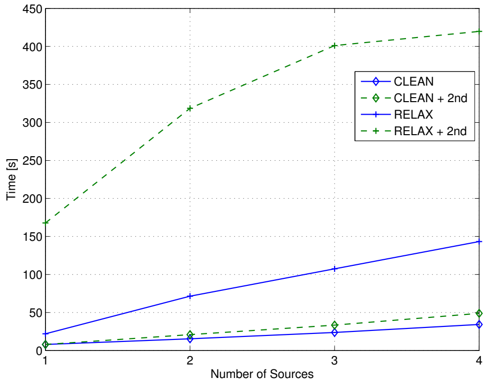


Figure 4.36: Computation time with respect number of sources

ber of loudspeaker available on the audio reproduction setup.

4.8.4 Additive noise on final configuration

Considering the previous results, the estimation performance of the methods is evaluated for different signal-to-noise ratios. For this experiment each of the four loudspeakers is used as source, while the other three are used to estimate the first-order reflections. The same procedure as in the previous part is used to obtain the final vertices positions estimates. The simulated RIRs have a reverberation time of $T60 = 0.2$s.

In Fig. 4.37 and Fig. 4.38 the average RMSE per vertex and its standard deviation, for each of the methods, are shown. From these results is seen that the RELAX-based methods offer the best performance. As it would be expected, the RMSE and it standard deviation decrease as the SNR increases. From all the methods the 2nd-Order RELAX seems to provide the best estimation performance. Even though RELAX has

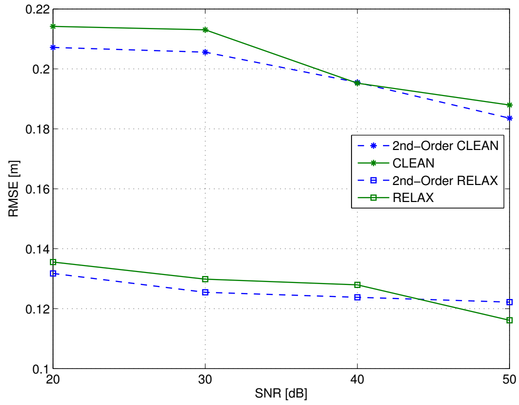


Figure 4.37: Average RMSE per vertex comparison considering loudspeaker transfer function

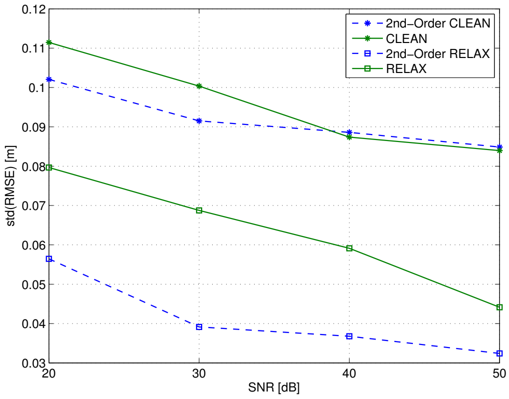


Figure 4.38: Standard deviation of error for the different methods

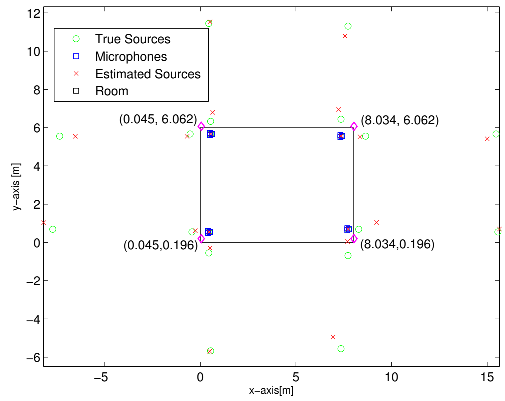


Figure 4.39: Estimation results from 2nd Order RELAX method

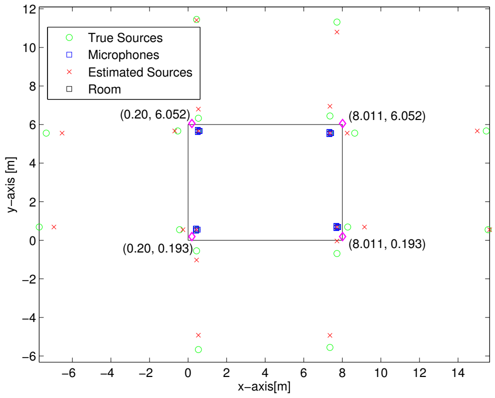


Figure 4.40: Estimation results from 2nd Order CLEAN method

a comparable performance to the 2nd-Order RELAX, its standard deviation is higher than the standard deviation of the latter. In Fig. 4.39 and Fig. 4.40 results of a single realization from two of the methods are shown. The estimated vertices are annotated in the images, and the true vertices are $\{[0,0],[8,0],[0,6],[8,6]\}$.

4.8.5 Comparison Wave-based Model vs EDM-based Model

Finally, here the performance of the methods discussed in this chapter, based on the propagation of waves, are compared with the ones discussed in Chapter 3, based on the properties of EDMs. In principle both representations are equivalent, i.e., through the TOA estimates a matrix can be constructed and used to estimate the locations of the image sources. However, while the methods based on EMDs solve a combinatorial problem, the methods presented in this chapter approach the problem using estimation theory. Basically, the iterative methods try to find the mixture of complex exponentials that produces the acquired data.

The results of this comparison are shown in Fig. 4.41 and Fig. 4.42. The configu-

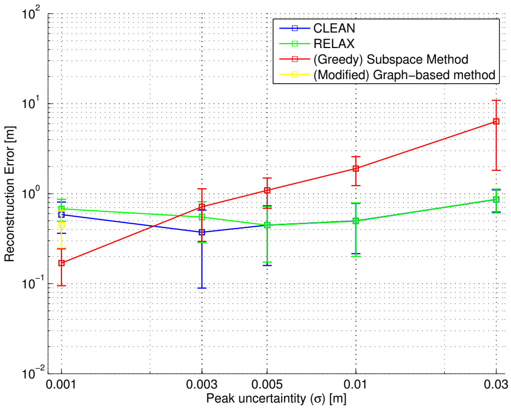


Figure 4.41: Room reconstruction error comparison for the different methods

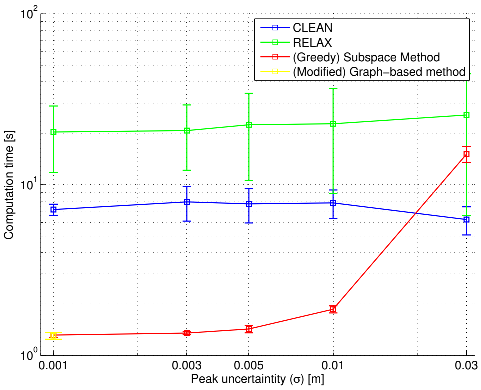


Figure 4.42: Computation time for each of the compared methods

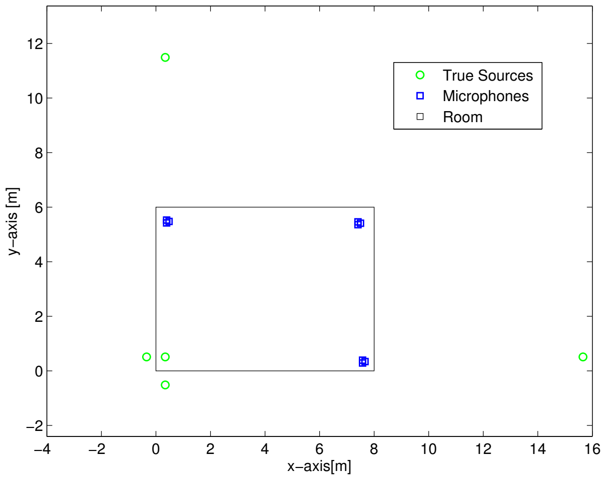


Figure 4.43: Configuration used for the comparison of the methods

ration used for the experiment is given in Fig. 4.43, where $M = 9$ microphones, three in each loudspeaker, are used to estimate the image sources generated by a fourth loudspeaker.

From these plots the effect of uncertainties in the TOA estimates in the estimation performance can be seen. Specially, the EDM-based methods are the most affected. In particular, the graph-based approach is not able to deliver results for most of the tested uncertainties due to the complexity of the maximum independent set listing problem. The high number of nodes of the graph render the NP-hard problem unfeasible to solve in reasonable time.

In the tested setup, the microphones are not distributed completely random, affecting the diversity of the TOA estimates. For this test, the microphones are allocated in UCAs with a radius of 6cm in each of the loudspeakers. This poses a difficulty for the EDM-based methods as the microphones positions are not sufficiently diverse to provide better results. Hence, they perform worse compared with the wave-based methods.

4.9 Discussion

In this chapter alternative methods, based on estimation theory, to estimate the first-order reflections were presented. The methods solve a high dimensional non-linear optimization problem by sequentially solving a set of two-dimensional optimization problems. It is shown that the estimator based on RELAX is a consistent and efficient estimator when there is complete agreement between the data and the assumed model. This property allows the estimators to outperform the methods based on EDMs when solving the problem of room geometry estimation for known TOA of first-order reflections. Furthermore, the performance of these estimators was evaluated when the data model is not met due to practical issues and the TOAs of the first-order reflections are not available. It was shown that it is possible to estimate the room geometry with a precision of circa 12cm in matter of several minutes. As all the results presented here consider simulated data, it should be noted that the estimation performance is expected to degrade even further when real measurements are used instead. These issues could be alleviated by increasing the number of sources and by only estimating first-order reflections within the directivity pattern of the driver, i.e., if a driver with cardiod directivity pattern is considered, the rear reflections should not be taken into consideration. Furthermore, in this work all the derivations are based on the assumption that the first-order reflections are the strongest contributions in our data model, however in real situations this is not necessary the case. Violations of this assumption further degrades the performance of the proposed methods in practice.

Conclusions 5

Inference of the room geometry is crucial for improvement of current methods for sound field estimation. Knowledge of the reproduction enclosure provides a natural way to address this problem by means of the parametrization of the RIRs. As a first step towards a general solution for sound field estimation in enclosures, we have presented two kind of methods capable to estimate the shape for shoe-box shaped rooms. These methods, closely related in principle, address the room geometry estimation problem from different perspectives.

In Chapter 3, it was shown that it is possible to solve the combinatorial problem of acoustic echoes labeling by means of a greedy strategy based on subspace techniques and the maximum rank property for EDMs. The proposed method shows comparable accuracy with respect to the current state-of-the-art method based on graph theory, at a reduced computational cost. Furthermore, it was shown that the devised subspace filtering can be used to further reduce the computational complexity of the graph-based approach. Under the assumptions established in Chapter 3, it was shown that the greedy method is able to estimate the vertices of the rooms with centimeters accuracy within seconds even in the presence of uncertainties in both TOA estimates and locations of the microphones.

Considering issues that could arise in practice, in Chapter 4 the room geometry estimation is done by posing the problem under an estimation theory framework. It was found that the efficient estimators developed in this chapter outperform the methods based on EDMs when the TOAs of each microphone-image source pair are known. Furthermore, the methods based on a mixture of complex exponential are more resilient to uncertainties in the TOA estimates compared to the ones based on EDMs. However, when we move away from the scenario of known RIRs, with identifiable first-order reflection peaks, and effects seen in a practical setup are included in the data, e.g., loudspeaker transfer function and reverberation, the estimators performance degrades considerably. In situations like these the proposed iterative methods provides estimates of the vertices of the rooms with an average error of 12cm within a couple of minutes.

5.1 Discussion

Through this thesis two different instances of the first-order reflections estimation problem for identifying the room geometry were considered. In each case different assumptions were made in order to shape the problems into relevant ones in both theory and practice. However, these assumptions might generate divided opinions leaving space for discussion. Hence, in this section a brief comment is made in order to clarify some of the decision made in this work.

In Chapter 3, the acoustic echoes sorting problem is discussed. In this chapter the existence of an *oracle* is assumed. This consideration can, perhaps, be the most arguable one. In most of real environments finding the peaks in the RIRs can be extremely hard. However, in some cases, when the early part of the RIRs is sparse enough, i.e., first-order reflections are perfectly identified, the sorting problem becomes relevant. In addition, the sorting problem is not only limited for acoustic echoes. The problem of sorting TOAs in a wireless sensor network could arise if there is no handshake between the *anchor* nodes and emerging nodes.

Besides the assumption of an oracle, through this thesis diversity in the microphones positions is considered. This assumption, as shown in Chapter 4, might not hold in standard audio reproduction setups, as it is assumed that the loudspeakers are found in a common horizontal plane. However, in home sound entertainment, despite existing recommendations for loudspeaker layouts, most of the users place the devices following aesthetics reasons. Hence, the loudspeakers, and as a result the microphones, most probably will be placed in arbitrary positions in the room, providing the necessary diversity for the methods to deliver appropriate results.

In Chapter 4, the estimation problem was treated as a joint estimation problem in the near field of a large array. However, it can be argued that the problem could have been tackled using individually each array under far field conditions. Under this approach, the UCA structure could have been exploited for fast computation of the cost function, i.e., for every DOA a fast Fourier transform can be employed to evaluate the range grid. However, the constraints in the number of microphones provide a poor angular resolution due to the width of the main lobe of the beam pattern. This situation hinders the ability to locate the first-order reflections by intersecting rays defined by the DOAs at each of the arrays.

5.2 Future Directions

The following ideas, which build upon the work presented in this thesis, could be useful for possible extensions and future research topics:

- **Sparse Acoustic Echoes Sorting**

  Due to my inclination towards convex optimization as a tool for approximating combinatorial problems, I would encourage the pursuit of a solution for the acoustic echoes sorting problem through sparse reconstruction and convex optimization. At this point, I can not provide any guarantee that such alternative exists, however the modeling of the problem as a (non-negative) matrix factorization problem could lead to further understanding of the underlying structure of the combinatorial problem. A possible model that could be explored is the following

  $$\mathbf{V} = \mathbf{P}\mathbf{S}_D^T\mathbf{R}_D, \tag{5.1}$$

  where $\mathbf{V} \in \mathbb{R}^{N\times M}$ is a permuted version of the transpose matrix of the true combinations of echoes $\mathbf{D} \in \mathbb{R}^{M\times N}$, $\mathbf{P} \in \mathbb{R}^{N\times(NM)}$ is a sparse matrix containing the concatenation of permutation matrices $\{\mathbf{P}_m \in \mathbb{R}^{N\times N}\}_{m=1}^{M}$, $\mathbf{S}_D \in \mathbb{R}^{(NM)\times(5M)}$ a block diagonal matrix which diagonal blocks are the structured $\mathbf{S} \in \mathbb{R}^{5\times N}$ matrix containing the positions of the (image) sources, and $\mathbf{R}_D \in \mathbb{R}^{\times 5M \times M}$ a sparse matrix with the known microphone positions.

- **Estimation outliers**
  While keeping the geometry of the room fixed to shoe-box shaped rooms, a natural extension of this work would be to deal with outliers in the iterative methods estimates. As the different image sources have distinct uncertainties, the image source with the lowest signal-to-interference-plus-noise ratio (SINR) usually is wrongly estimated. This heavily biases the vertices estimates degrading the overall performance of the approach. This issue becomes of great importance when dealing with directional loudspeakers, i.e., the rear first-order reflection is heavily attenuated due to the directivity of the loudspeaker.

- **Exploiting loudspeaker directivity**
  Considering that the loudspeaker directivity can be known beforehand, i.e., as product manufacturer the full characterization of the loudspeaker is known, it is possible to add this knowledge to the iterative methods in Chapter 4 in order to try to search for further improvements in practical performance.

- **Non-convex rooms**
  Even though the work in this thesis was focused in shoe-box shaped rooms, the methods presented here can be extended to any arbitrary room shape. The extension to non-convex boundaries could be possible by *mapping* the room at different locations. This is possible as any room boundary can be described by a piece-wise continuous function. A possible way to reconstruct the boundary of a non-convex room could require a moving source. This is done in order to find the reflex interior angles of the polygon describing the non-convex shape.

- **A different representation**

  Through this work a fundamental assumption was made: there is no *intuitive* structure in the RIRs. However, this might be true for RIRs defined as FIR filters based on finite number of taps but not for other type of models, such as the ones based on orthonormal basis functions (OBFs) [48][47]. Hence, it would be of interest to explore the possibility to relate this kind of room modeling to the room geometry estimation problem, as these models could offer a different approach to RIRs re-parametrization in terms of common room resonances.

Euclidean Distance Matrices A

This appendix contains the basic theory behind Euclidean distance matrices. For more information of this topic and its applications, the reader is referred to [11].

A.1 Generalities

An **Euclidean Distance Matrix (EDM)** $\mathbf{D} \in \mathbb{R}^{M\times M}$ is a matrix of squared Euclidean distances between a set of M points in a N-dimensional Euclidean space. Consider a set $\mathcal{X} = \{\mathbf{x}_m \in \mathbb{R}^N\}_{m=1}^M$, ascribed to the columns of the matrix $\mathbf{X} \in \mathbb{R}^{N\times M}$, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_M]$. The entries of $\mathbf{D} \in \mathbb{EDM}$ are the squared Euclidean distances between the (i, j)-th pair of element given by

$$[\mathbf{D}]_{ij} = d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|^2, \tag{A.1}$$

where $\|\cdot\|$ is the Euclidean norm. Expanding the previous expression yields

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

From (A.2) the matrix equation for the elements in $\mathbf{D}$ can be expressed as

$$edm(\mathbf{X}) \triangleq \mathbf{D} = \mathbf{1}\text{diag}(\mathbf{X}^T\mathbf{X})^T - 2\mathbf{X}^T\mathbf{X} + \text{diag}(\mathbf{X}^T\mathbf{X})\mathbf{1}^T, \tag{A.3}$$

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

Theorem 1. *(Rank of EDM) The rank of $\boldsymbol{D} \in \mathbb{EDM}$ corresponding to points in a N-dimensional space is at most $N + 2$.*

Proof. Observe that $rank(\mathbf{X}^T\mathbf{X}) \leq N$ as $rank(\mathbf{X}) \leq N$ and that the other two terms has rank one. Using the rank inequality for sum of matrices the following is obtained

$$\begin{aligned} rank(\mathbf{D}) &\leq rank(\mathbf{1}\text{diag}(\mathbf{X}^T\mathbf{X})^T) + rank(2\mathbf{X}^T\mathbf{X}) + rank(\text{diag}(\mathbf{X}^T\mathbf{X})\mathbf{1}^T) \\ &\leq N + 2 \end{aligned} \tag{A.4}$$

□

Definition 1. *(Affine Dimension of EDM) For a matrix $\boldsymbol{D} \in \mathbb{EDM}$, its embedding or affine dimension affdim($\boldsymbol{D}$) is the dimension of the smallest subspace that contains points capable of generate $\boldsymbol{D}$.*

The definition of affdim($\mathbf{D}$) provides an insight on how the points are structure in the space. Consider the set $\mathcal{X}_2$ of 2D points distributed along a line. The EDM generated by these points, is also generate by a set $\mathcal{X}_1$ which contains 1D points maintaining the same distances as in the 2D case. Hence, it can be concluded that are infinitely many points sets able to generate a given EDM.

Theorem 2. *(Rigid Transformation Invariance of EDMs) The set of points $\mathcal{X}$ of dimension equal to the affdim($\boldsymbol{D}$) that generates $\boldsymbol{D}$ can only be reconstructed from $\boldsymbol{D}$ upto a rigid transformation.*

Proof. First notice that $\mathbf{D}$ is a function of $\mathbf{X}^T\mathbf{X} \in \mathbb{R}^{M\times M}$. Then, consider any rotation/reflection acting over the points $X \in \mathbb{R}^{N\times M}$, represented by an orthogonal matrix $\mathbf{Q} \in \mathbb{R}^{N\times N}$. The Gram matrix of the rotated/reflected points $\mathbf{X}_r = \mathbf{QX}$ is given by

$$\mathbf{X}_r^T\mathbf{X}_r = \mathbf{X}^T\mathbf{Q}^T\mathbf{Q}\mathbf{X}^T = \mathbf{X}^T\mathbf{X} \tag{A.5}$$

where the fact that $\mathbf{Q}$ is an orthogonal matrix, i.e., $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$ has been used. This proves that rotations and reflections does not alter the distances in $\mathbf{D}$. Now, consider a translation of the points, i.e.,

$$\mathbf{X}_t = \mathbf{X} + \mathbf{b}\mathbf{1}^T, \ \ \mathbf{b} \in \mathbb{R}^N. \tag{A.6}$$

By observing that $\operatorname{diag}(\mathbf{X}_t^T\mathbf{X}_t) = \operatorname{diag}(\mathbf{X}^T\mathbf{X}) + 2\mathbf{X}^T\mathbf{b} + \mathbf{b}^T\mathbf{b}\mathbf{1}$, it can be verified that this translation does not make any changes in (A.4). This proves its invariance with respect translations. Finally, the following result can be stated

$$edm(\mathbf{QX}) = edm(\mathbf{X} + \mathbf{b}\mathbf{1}^T) = edm(\mathbf{X}) \tag{A.7}$$

□

A direct consequence of this theorem is the impossibility to reconstruct the absolute coordinate of the generating points. Every distinct reconstruction procedure to retrieve $\mathbf{X}$ from $\mathbf{D}$ leads to different a realization of the set of points only differing by a rigid transform.

Finally, let us introduce the last theorem that provides the necessary and sufficient conditions for an arbitrary matrix to be an EDM. In order to state the theorem, two definitions should be presented before.

Definition 2. *(Symmetric hollow subspace) Denoted by $\mathcal{S}_H^N$, the symmetric hollow subspace is a proper subspace of symmetric matrices $\mathcal{S}^N$ with zero diagonal.*

$$\mathcal{S}_H^N \overset{def}{=} \{\boldsymbol{A} \in \mathcal{S}^N \mid diag(\boldsymbol{A}) = \boldsymbol{0}\}. \tag{A.8}$$

Definition 3. *(Positive semi-definite cone) Denoted by $\mathcal{S}_+^N$, the positive semi-definite cone is the set of all symmetric positive semi-definite matrices of dimensions $N \times N$*

$$\mathcal{S}_+^N \overset{def}{=} \{\boldsymbol{A} \in \mathcal{S}^N \mid \boldsymbol{A} \succeq \boldsymbol{0}\}. \tag{A.9}$$

Theorem 3. *(GOWER [18]) Let be a geometric centering matrix be given by*

$$\boldsymbol{L} \overset{def}{=} \boldsymbol{I} - \frac{1}{N}\boldsymbol{1}\boldsymbol{1}^T, \tag{A.10}$$

where $\boldsymbol{I}_N$ is a $N \times N$ identity matrix. Then,

$$\boldsymbol{D} \in \mathbb{EDM} \iff \begin{cases} -\boldsymbol{LDL} \in \mathcal{S}_+^N \\ \boldsymbol{D} \in \mathcal{S}_H^N \end{cases} \tag{A.11}$$

Proof. Found in reference [18]. □

Graph Theory B

In this appendix a brief introduction to the concepts of graph theory needed for this thesis is presented. This is not, by any means, an in-depth discussion of graph theory, for a more thorough treatment of these concepts the reader is referred to [9].

B.1 General Definitions

In general, a **graph** is formed by *vertices* (nodes) and *edges* connecting the vertices.

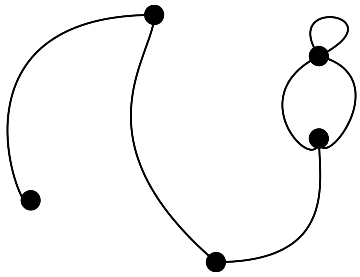


Figure B.1: Graph $G(V, E)$ with $|V(G)| = 5$ and $|E(G)| = 6$

Formally, a **graph** $G = (V, E)$ is defined as a pair consisting of a **vertex set** $V(G)$, an **edge set** $E(G)$, and a relation that associates with each edge, two vertices called its **endpoints**. When a graph does not distinguish between the ordering of the endpoints, it is known as **undirected graph**, otherwise it is considered as a **directed graph**. In practice, a vertex can be used to represent anything, e.g., candidates of sorted distances, and edges are used to indicate relations between vertices, e.g., candidates sharing elements in common. Fig. B.1 shows a typical graphical representation of a graph.

A **simple graph** is a graph having no loops or multiple edges. This means that there is no more than one edge sharing the same endpoints and that the endpoints of an edge are not the same vertex. In a simple graph each edge $e \in E(G)$ can be uniquely identified by its endpoints $u, v \in V(G)$. Two vertices $u, v \in V(G)$ are considered **adjacent** if they define an edge in the graph as $e = uv$. A simple graph that contains every possible edge between all the vertices is called a **complete graph**.

B.2 Independent Sets and Cliques

A set of vertices $V'(G) \subset V(G)$ that are not pairwise adjacent is known as **independent set**. In Fig. B.2 the subset $V'(G) = \{1, 3, 7\}$ is an independent set, as these three vertices does not share any edge in common. In addition, all the subsets of an independent set are also considered independent sets, e.g. $\tilde{V}(G) = \{1, 7\} \subset V'(G)$. When

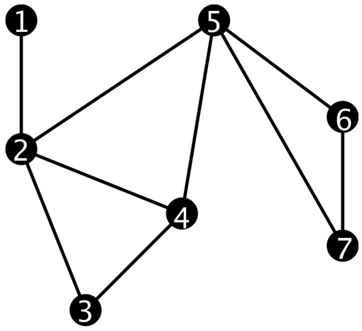


Figure B.2: Simple graph $G(V, E)$. Neither loops or multiple edges present.

a set $\tilde{V}(G)$ cannot include any other vertex without forcing it to have an edge, the set is known as **maximal independent set**. The maximal independent sets with the largest allowable cardinality in the graph are called **maximum independent sets**. An example of these types of sets is shown in Fig. B.3, where three different sets denoted by blue vertices are selected from the same graph.

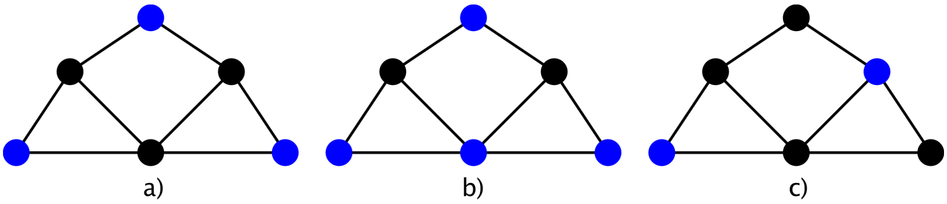


Figure B.3: Sets taken from a given graph $G(V, E)$. a) Maximum independent set, b) Not independent set, c) Maximal independent set

A **clique** can be seen as a complementary definition of an independent set, as it is a set of pairwise adjacent vertices. In other words, a clique is a set of vertices which induced graph is a complete graph. Analogously to the independent sets, a **maximal clique** can be defined as a clique that cannot be extended by including one more adjacent vertex. A **maximum clique** is a maximal clique with the largest allowable cardinality. Fig. B.4 exemplifies the clique concepts by taking three different vertices subsets denoted by blue nodes.

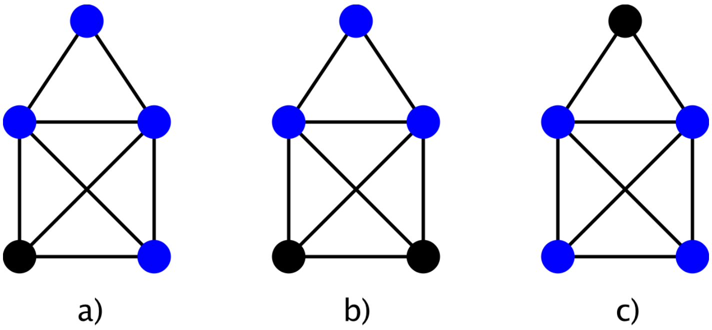


Figure B.4: Sets taken from a given graph $G(V, E)$. a) Not a clique, b) Maximal clique, c) Maximum clique

B.3 Complement of a Graph

The **complement** or inverse of a graph G is a graph H on the same vertices such that two different vertices in H are adjacent if and only if they are not adjacent in G. In order to build H from G it is only needed to add the missing edges until G becomes complete, and then remove the original edges existing in G. By using the complementary nature of cliques and independent sets, it is seen that any independent set in the graph G is a clique in the graph H and vice versa. A graph G and its complement H are shown in Fig. B.5, where the set of blue vertices represent a clique and independent set in the original and complement graph respectively.

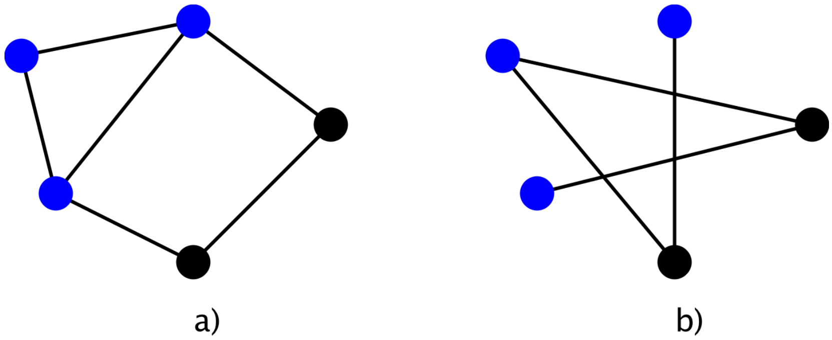


Figure B.5: a) Graph G, b) Graph H : Complement of graph G

Due to this complementary property of the independents sets and cliques, the maximal independent set listing problem can be cast as a maximal clique listing problem instead. By transforming the original graph into its complement H, algorithms that finds all maximum cliques can be employed to solve the echo disambiguation problem presented in this thesis.

Estimation Theory C

This appendix contains a brief introduction to estimation theory. For more information of this topic and its applications the reader is referred to [27].

C.1 Generalities

The field of estimation theory deals with estimating the values of unknown parameters based on measured data which contains a random component. Using the assumption that the unknown parameters interact with the measurements by modifying their distribution, it is possible to design estimators capable to approximate these parameters from the available data.

For example, consider the temperature of a room being measured. The intrinsic noise of the employed thermal sensor randomly distributes the measured values, so that the true temperature must be estimated. If the noise is not random, the problem could be considered deterministic, i.e., removing an offset, and estimation would not be necessary.

In order to apply estimation theory to a given problem, it is required to have knowledge of the *measured* data $\mathbf{x}$ with $\dim(\mathbf{x}) = N$, and the *model* that describes the interaction between the data and the unknown parameters $\boldsymbol{\theta}$ with $\dim(\boldsymbol{\theta}) = M$. To illustrate this, consider again the example of measuring the temperature of a room. The sampled measured signal from the thermal sensor can be modeled as

$$\mathbf{x}[n] = \theta + w[n] \in \mathbb{R},\ n = 0, \dots, N-1, \tag{C.1}$$

where θ is the true temperature and $w[n]$ is considered to be additive zero-mean white Gaussian noise with variance σ^2, i.e., $p(w[n]) \sim \mathcal{N}(0, \sigma^2)$, representing the sensor noise. By using the model in C.1 it is possible to express the likelihood function, defining how the parameter θ affects the distribution of the data $\mathbf{x}$, by using the original distribution of the noise. That is, if

$$p(w[n]) \sim \mathcal{N}(0, \sigma^2), \tag{C.2}$$

the measured data would be distributed accordingly with

$$p(x[n]) \sim \mathcal{N}(\theta, \sigma^2), \tag{C.3}$$

and the likelihood function $p(x[n]; \theta)$ would be given by

$$p(x[n]; \theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\big(- \frac{1}{2\sigma^2} (x[n] - \theta)^2 \big). \tag{C.4}$$

The expression in (C.4) shows how the unknown parameter θ affects the distribution of our measured data. The goal of estimation theory is the estimation of θ, given that there is knowledge of measured data $\{\mathbf{x}[n]\}_0^{N-1}$.

One of most widely used estimators, is the one known as the maximum likelihood estimator (MLE) [27]. This estimator selects the parameter θ which maximizes the likelihood function. Intuitively, the MLE maximizes the agreement of the observed data with the employed model. This kind of estimator, due to its properties of consistency and efficiency [27], is used as foundation for the estimators proposed in Chapter 4.

C.2 Fisher Information and Cramér-Rao Lower Bound

C.2.1 Fisher Information

In general terms, the Fisher information (FI) is a mathematical descriptor of the amount of information a measurement $\mathbf{x}$ contains about a unknown parameter vector $\boldsymbol{\theta}$ that contributes to its statistical behavior.

By defining the likelihood function of $\mathbf{x}$ due to the parameter $\boldsymbol{\theta}$ as $p(\mathbf{x}|\theta)$, the elements of the Fisher information matrix (FIM) $\mathbf{J}(\boldsymbol{\theta})$ are given by

$$\mathbf{J}_{i,j}(\boldsymbol{\theta}) = -E\{\frac{\partial^2 \ln p(\mathbf{x}|\boldsymbol{\theta}_i)}{\partial\boldsymbol{\theta}_i\partial\boldsymbol{\theta}_j}\}. \tag{C.5}$$

From a geometrical point of view, considering the log likelihood function as a *score* that indicates how sensitive $p(\mathbf{x}|\boldsymbol{\theta})$ is to $\boldsymbol{\theta}$ and assuming that certain regularity conditions hold, the FIM represents the curvature of the support near the MLE of $\boldsymbol{\theta}$. This can be mathematically expressed as the covariance of the score function $\frac{\partial \ln p(\mathbf{x}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_i}$ given by

$$\mathbf{J}_{i,j}(\boldsymbol{\theta}) = E\{\frac{\partial \ln p(\mathbf{x}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_i}\frac{\partial \ln p(\mathbf{x}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_j}\}. \tag{C.6}$$

Intuitively, a sharp peak around the MLE has a *high* curvature, hence high information is carried by $\mathbf{x}$. On the other hand, a flat log likehood function reveals a small curvature rendering the variate insensitive to $\boldsymbol{\theta}$.

C.2.2 Cramér Rao Lower Bound

H. Cramer and C.R Rao introduced this bound as the lowest attainable variance [27] of any unbiased estimator $\hat{\boldsymbol{\theta}}$, i.e.,

$$\text{Bias}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = (E\{\hat{\boldsymbol{\theta}}\} - \boldsymbol{\theta})^2 = 0. \tag{C.7}$$

The Cramér Rao lower bound (CRLB) is given by

$$\text{Var}(\hat{\boldsymbol{\theta}}) \geq CRLB(\boldsymbol{\theta}) = \mathbf{J}^{-1}(\boldsymbol{\theta}), \tag{C.8}$$

which is nothing more than the inverse of the FIM. Hence, the mean square error,

$$\text{MSE}(\hat{\boldsymbol{\theta}}) = E\{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^2\}, \tag{C.9}$$

for an efficient unbiased estimator, i.e., unbiased estimator achieving the CRLB, is given by

$$\begin{aligned} \text{MSE}(\hat{\boldsymbol{\theta}}) &= \text{Bias}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})^2 + \text{Var}(\hat{\boldsymbol{\theta}}) && \text{(C.10)} \\ &= \text{Var}(\hat{\boldsymbol{\theta}}) = CRLB(\boldsymbol{\theta}). && \text{(C.11)} \end{aligned}$$

Cramér-Rao Lower Bound for Source Localization D

In this appendix the Cramér-Rao lower bound (CRLB) for source localization with known deterministic source signal is derived. The bounds for both near field and far field cases are presented.

D.1 Near field - Single Source CRLB

Consider the following model

$$\mathbf{y}(\omega) = \mathbf{a}(\mathbf{x}, \omega)s(\omega) + \mathbf{n}(\omega) \in \mathbb{C}^{M\times 1}, \tag{D.1}$$

where $\mathbf{x} \in \mathbb{R}^D$ denotes the unknown vector parameter, containing the position of the source in a D dimensional space. Assuming that

$$\begin{aligned} E\{\mathbf{y}(\omega)\} &= \mathbf{a}(\mathbf{x}, \omega)s(\omega), && \text{(D.2)} \\ \mathbf{R}_{yy}(\omega) &= \sigma_N^2\mathbf{I}, && \text{(D.3)} \end{aligned}$$

the Fisher information, for a single frequency ω, is given by[1]

$$\begin{aligned} J(\mathbf{x}, \omega) &= \frac{2}{\sigma_N^2}\mathrm{Re}\left\{\left(\frac{\partial\mathbf{a}(\mathbf{x}, \omega)}{\partial\mathbf{x}}\right)^H\left(\frac{\partial\mathbf{a}(\mathbf{x}, \omega)}{\partial\mathbf{x}}\right)\right\}\|s(\omega)\|^2 && \text{(D.4)} \\ &= 2SNR(\omega)\mathrm{Re}\left\{\left(\frac{\partial\mathbf{a}(\mathbf{x}, \omega)}{\partial\mathbf{x}}\right)^H\left(\frac{\partial\mathbf{a}(\mathbf{x}, \omega)}{\partial\mathbf{x}}\right)\right\} \in \mathbb{R}^{D\times D}. && \text{(D.5)} \end{aligned}$$

For the near field case, the m-th element of the array vector response is given by

$$[\mathbf{a}(\mathbf{x}, \omega)]_m = \frac{1}{r_m(\mathbf{x})}\exp(-j\omega r_m(\mathbf{x})/c) \in \mathbb{C}. \tag{D.6}$$

where c is the speed of sound and $r_m(\mathbf{x}) = \|\mathbf{x} - \mathbf{p}_m\|$ is the distance between the source $\mathbf{x}$ and the m-th array element at coordinate $\mathbf{p}_m \in \mathbb{R}^D$.

To provide a close form for the expression in (D.5), the m-th row of the steering vector gradient matrix is found to be given by

$$\left[\frac{\partial}{\partial\mathbf{x}}\mathbf{a}(\mathbf{x}, \omega)\right]_m = \left(\frac{\partial}{\partial\mathbf{x}}r_m^{-1}(\mathbf{x})\right)\exp(-j\omega r_m(\mathbf{x})/c) + r_m^{-1}(\mathbf{x})\left(\frac{\partial}{\partial\mathbf{x}}\exp(-j\omega r_m(\mathbf{x})/c)\right) \in \mathbb{C}^{1\times D}, \tag{D.7}$$

where

$$\frac{\partial}{\partial\mathbf{x}}r_m^{-1}(\mathbf{x}) = \frac{\mathbf{p}_m - \mathbf{x}}{\|\mathbf{x} - \mathbf{p}_m\|^3} \in \mathbb{C}^{1\times D}, \tag{D.8}$$

[1] Chen, et al ”A maximum-likelihood parametric approach to source localizations.” ICAASP 2001

$$\frac{\partial}{\partial \mathbf{x}}\exp(-j\omega r_m(\mathbf{x})/c) = -j\omega/c\bigg(\frac{\mathbf{x}-\mathbf{p}_m}{\|\mathbf{x}-\mathbf{p}_m\|}\bigg)\exp(-j\omega r_m(\mathbf{x})/c) \in \mathbb{C}^{1\times D}, \tag{D.9}$$

$\therefore$

$$\left[\frac{\partial}{\partial \mathbf{x}}\mathbf{a}(\mathbf{x},\omega)\right]_m = \frac{\mathbf{p}_m-\mathbf{x}}{\|\mathbf{x}-\mathbf{p}_m\|^2}\exp(-j\omega r_m(\mathbf{x})/c)\left[\frac{1}{\|\mathbf{x}-\mathbf{p}_m\|}+j\omega/c\right] \in \mathbb{C}^{1\times D}. \tag{D.10}$$

Now, the (i,j)-th entry of the gradient product matrix is found to be

$$\left[\left(\frac{\partial \mathbf{a}(\mathbf{x},\omega)}{\partial \mathbf{x}}\right)^H\left(\frac{\partial \mathbf{a}(\mathbf{x},\omega)}{\partial \mathbf{x}}\right)\right]_{ij} = \sum_{m=1}^{M}\frac{[\mathbf{p}_m-\mathbf{x}]_i[\mathbf{p}_m-\mathbf{x}]_j}{\|\mathbf{x}-\mathbf{p}_m\|^4}\bigg(\frac{1}{\|\mathbf{x}-\mathbf{p}_m\|^2}+\frac{\omega^2}{c^2}\bigg). \tag{D.11}$$

Hence, the (i,j)-th entry of the Fisher information for a single frequency ω can be expressed as

$$\big[J(\mathbf{x},\omega)\big]_{ij} = 2SNR(\omega)\left[\sum_{m=1}^{M}\frac{[\mathbf{p}_m-\mathbf{x}]_i[\mathbf{p}_m-\mathbf{x}]_j}{\|\mathbf{x}-\mathbf{p}_m\|^4}\bigg(\frac{1}{\|\mathbf{x}-\mathbf{p}_m\|^2}+\frac{\omega^2}{c^2}\bigg)\right]. \tag{D.12}$$

Assuming constant power density for the signal and the noise inside a certain frequency band $[\omega_c - B/2, \omega_c + B/2]$, and some processing window of length T, the (i,j)-th entry of the Fisher information over all the frequency band can be calculated as

$$\big[J(\mathbf{x})\big]_{ij} = \frac{2T}{2\pi}SNR\int_{\omega_c-B/2}^{\omega_c+B/2}\left[\sum_{m=1}^{M}\frac{[\mathbf{p}_m-\mathbf{x}]_i[\mathbf{p}_m-\mathbf{x}]_j}{\|\mathbf{x}-\mathbf{p}_m\|^4}\bigg(\frac{1}{\|\mathbf{x}-\mathbf{p}_m\|^2}+\frac{\omega^2}{c^2}\bigg)\right]d\omega, \tag{D.13}$$

leading to

$$\big[J(\mathbf{x})\big]_{ij} = \frac{T}{\pi}SNR\bigg[B\sum_{m=1}^{M}\frac{[\mathbf{p}_m-\mathbf{x}]_i[\mathbf{p}_m-\mathbf{x}]_j}{\|\mathbf{x}-\mathbf{p}_m\|^6}+ \\ (B\omega_c^2+B^3/12)c^{-2}\sum_{m=1}^{M}\frac{[\mathbf{p}_m-\mathbf{x}]_i[\mathbf{p}_m-\mathbf{x}]_j}{\|\mathbf{x}-\mathbf{p}_m\|^4}\bigg]. \tag{D.14}$$

Finally, the CRLB for the parameter $\mathbf{x}$ is given by

$$CRLB(\mathbf{x}) = J(\mathbf{x})^{-1}. \tag{D.15}$$

D.2 Far field - Single Source CRLB

For the far field case the model in (D.1) and the assumptions in (D.2) and (D.3) are considered. Then, the Fisher information matrix has the same form of (D.5) given by

$$J(\mathbf{x},\omega) = 2SNR(\omega)\bigg(\frac{\partial \mathbf{a}(\mathbf{x},\omega)}{\partial \mathbf{x}}\bigg)^H\bigg(\frac{\partial \mathbf{a}(\mathbf{x},\omega)}{\partial \mathbf{x}}\bigg) \in \mathbb{R}^{D\times D}. \tag{D.16}$$

However, in the far-field case, the i-th element of the array vector response is given instead by

$$[\mathbf{a}(\mathbf{x}, \omega)]_m = \exp(-j\omega r_m(\mathbf{x})/c). \tag{D.17}$$

Following a similar procedure as in the near-field case, the Fisher information over all the frequency band $[\omega_c - B/2, \omega_c + B/2]$ is found to be[2]

$$J(\mathbf{x}) = \frac{T}{\pi} SNR(B\omega_c^2 + B^3/12) \left[\left(\partial \mathbf{R}\right)^T \left(\partial \mathbf{R}\right) \right] \in \mathbb{R}^{D \times D}, \tag{D.18}$$

with the following definition

$$\partial \mathbf{R} \triangleq \frac{1}{c} \begin{bmatrix} \frac{\partial r_1(\mathbf{x})}{\partial \mathbf{x}} \\ \vdots \\ \frac{\partial r_M(\mathbf{x})}{\partial \mathbf{x}} \end{bmatrix} \in \mathbb{R}^{M \times D},$$

where

$$\frac{\partial r_m(\mathbf{x})}{\partial \mathbf{x}} = \frac{\mathbf{x} - \mathbf{p}_m}{\|\mathbf{x} - \mathbf{p}_m\|}.$$

D.3 Near Field - Multiple Source CRLB

In the case that more than one source is present, e.g., direct path and first order reflections, the CRLB for more than one source in the near field is derived.

Consider the following model for a single frequency

$$\mathbf{y}(\omega) = \left[\sum_{n=1}^{N} \mathbf{a}(\mathbf{x}_n, \omega) \right] s(\omega) + \mathbf{n}(\omega). \tag{D.19}$$

Using similar assumptions as in the single source case, the following signal's statistics are considered in the model

$$E\{\mathbf{y}(\omega)\} = \left[\sum_{n=1}^{N} \mathbf{a}(\mathbf{x}_n, \omega) \right] s(\omega), \tag{D.20}$$

$$\mathbf{R}_{yy}(\omega) = \sigma_N^2 \mathbf{I}. \tag{D.21}$$

The Fisher information for the multiple source case given in (D.22) is a block matrix whose diagonal elements are the single source Fisher matrices for each source, i.e.,

$$J(\mathbf{x}, \omega) = \begin{bmatrix} J_{11}(\mathbf{x}, \omega) & \cdots & J_{1N}(\mathbf{x}, \omega) \\ \vdots & \ddots & \vdots \\ J_{N1}(\mathbf{x}, \omega) & \cdots & J_{NN}(\mathbf{x}, \omega) \end{bmatrix} \in \mathbb{R}^{ND \times ND}, \tag{D.22}$$

[2] Similar expression as in: Tervo, Sakari. *"Localization and tracing of early acoustic reflections."* (2011).

where $\mathbf{x} = [\mathbf{x}_1^T, \ldots, \mathbf{x}_N^T]^T \in \mathbb{R}^{ND}$ is the unknown parameter vector containing the positions of the sources $\mathbf{x}_i \in \mathbb{R}^D \ \forall \ i \in \{1, \ldots, N\}$. The (i,j)-th block from $J(\mathbf{x}, \omega)$ is given by

$$\begin{aligned} J_{ij}(\mathbf{x}, \omega) &= \frac{2\|s(\omega)\|^2}{\sigma_N^2} \mathrm{Re}\bigg\{ \Big(\frac{\partial}{\partial \mathbf{x}_i} \sum_{n=1}^{N} \mathbf{a}(\mathbf{x}_n, \omega) \Big)^H \Big(\frac{\partial}{\partial \mathbf{x}_j} \sum_{n=1}^{N} \mathbf{a}(\mathbf{x}_n, \omega) \Big) \bigg\} && \text{(D.23)} \\ &= 2SNR(\omega) \mathrm{Re}\bigg\{ \Big(\frac{\partial}{\partial \mathbf{x}_i} \mathbf{a}(\mathbf{x}_i, \omega) \Big)^H \Big(\frac{\partial}{\partial \mathbf{x}_j} \mathbf{a}(\mathbf{x}_j, \omega) \Big) \bigg\} \in \mathbb{R}^{D \times D}. && \text{(D.24)} \end{aligned}$$

As in the single source, the gradient of the steering vector, for the m-th array element, is given by

$$\left[\frac{\partial}{\partial \mathbf{x}_i} \mathbf{a}(\mathbf{x}_i, \omega) \right]_m = \frac{\mathbf{p}_m - \mathbf{x}_i}{\|\mathbf{x} - \mathbf{p}_m\|^2} \exp(-j\omega r_m(\mathbf{x}_i)/c) \left[\frac{1}{\|\mathbf{x}_i - \mathbf{p}_m\|^2} + j\omega/c \right] \in \mathbb{C}^{1 \times D}. \quad \text{(D.25)}$$

Hence, the (k,l)-th entry of the (i,j)-th block of $J(\mathbf{x}, \omega)$ can be expressed as

$$\begin{aligned} \big[J_{ij}(\mathbf{x}, \omega) \big]_{kl} = 2SNR(\omega) \mathrm{Re}\bigg\{ \sum_{m=1}^{M} & \frac{[\mathbf{p}_m - \mathbf{x}_i]_k}{\|\mathbf{x}_i - \mathbf{p}_m\|^2} \exp(j\omega r_m(\mathbf{x}_i)/c) \left[\frac{1}{\|\mathbf{x}_i - \mathbf{p}_m\|} - j\omega/c \right] \times \\ & \frac{[\mathbf{p}_m - \mathbf{x}_j]_k}{\|\mathbf{x}_j - \mathbf{p}_m\|^2} \exp(-j\omega r_m(\mathbf{x}_j)/c) \left[\frac{1}{\|\mathbf{x}_j - \mathbf{p}_m\|} + j\omega/c \right] \bigg\}. \quad \text{(D.26)} \end{aligned}$$

Collecting terms and introducing new variables, the condensed expression for $\big[J_{ij}(\mathbf{x}, \omega) \big]_{kl}$ is given by

$$\big[J_{ij}(\mathbf{x}, \omega) \big]_{kl} = 2SNR(\omega) \mathrm{Re}\bigg\{ \sum_{m=1}^{M} A_{ij}^{(m)}(k,l) \exp(j\omega \alpha_{ij}^{(m)}) \Big(B_{ij}^{(m)} + j\frac{\omega}{c} C_{i,j}^{(m)} + \frac{\omega^2}{c^2} \Big) \bigg\}, \quad \text{(D.27)}$$

where the following definitions has been made

$$\begin{aligned} \alpha_{ij}^{(m)} &\triangleq c^{-1}\big(r_m(\mathbf{x}_i) - r_m(\mathbf{x}_j) \big), && \text{(D.28)} \\ A_{ij}^{(m)}(k,l) &\triangleq \frac{[\mathbf{p}_m - \mathbf{x}_i]_k [\mathbf{p}_m - \mathbf{x}_j]_l}{\|\mathbf{x}_i - \mathbf{p}_m\|^2 \|\mathbf{x}_j - \mathbf{p}_m\|^2}, && \text{(D.29)} \\ B_{ij}^{(m)} &\triangleq \frac{1}{\|\mathbf{x}_i - \mathbf{p}_m\| \|\mathbf{x}_j - \mathbf{p}_m\|}, && \text{(D.30)} \\ C_{ij}^{(m)} &\triangleq \frac{1}{\|\mathbf{x}_i - \mathbf{p}_m\|} - \frac{1}{\|\mathbf{x}_j - \mathbf{p}_m\|}. && \text{(D.31)} \end{aligned}$$

Using Euler's formula to express the complex exponential in terms of its real and imaginary terms, (D.27) can be expressed as

$$\big[J_{ij}(\mathbf{x}, \omega) \big]_{kl} = 2SNR(\omega) \sum_{m=1}^{M} A_{ij}^{(m)}(k,l) \left[\cos(\omega \alpha_{ij}^{(m)}) \big(B_{ij}^{(m)} + \frac{\omega^2}{c^2} \big) - \sin(\omega \alpha_{ij}^{(m)}) C_{ij}^{(m)} \frac{\omega}{c} \right]. \quad \text{(D.32)}$$

From expression in (D.32) the single source, i.e., $i = j$, Fisher information is readily derivable. For the diagonal blocks of the multiple sources Fisher information the equality $\alpha_{ii}^{(m)} = 0 \; \forall \; i, m$ holds. Hence, these elements can be obtained directly from (D.13).

Similarly as in the previous CRLBs, the power density for the signal and noise inside the frequency band $[\omega_c - B/2, \omega_c + B/2]$ is assumed constant, and a processing window of length T is considered. As a result, the (k, l)-th element from the (i, j)-th block of the Fisher information over all the bandwidth $\forall \; i \neq j$ is given by

$$\big[J_{ij}(\mathbf{x})\big]_{kl} = \frac{T}{2\pi} \int\limits_{\omega_c - B/2}^{\omega_c + B/2} \Big[J_{ij}(\mathbf{x}, \omega)\Big]_{kl} d\omega. \tag{D.33}$$

Substituting (D.32) in (D.33)

$$\begin{aligned}\big[J_{ij}(\mathbf{x})\big]_{kl} = \frac{T}{\pi} SNR \sum_{m=1}^{M} A_{ij}^{(m)}(k,l) \Bigg[B_{ij}^{(m)} \int\limits_{\omega_c - B/2}^{\omega_c + B/2} \cos(\omega \alpha_{ij}^{(m)}) d\omega + \\ c^{-2} \int\limits_{\omega_c - B/2}^{\omega_c + B/2} \omega^2 \cos(\omega \alpha_{ij}^{(m)} d\omega) - c^{-1} C_{ij}^{(m)} \int\limits_{\omega_c - B/2}^{\omega_c + B/2} \omega \sin(\omega \alpha_{ij}^{(m)}) d\omega \Bigg].\end{aligned} \tag{D.34}$$

By solving the integrals in (D.34), the following expression can be evaluated to find the value of $\big[J_{ij}\big]_{kl} \; \forall \; i \neq j$

$$\begin{aligned}\big[J_{ij}(\mathbf{x}, \omega)\big]_{kl} = \frac{T}{\pi} SNR \sum_{m=1}^{M} A_{ij}^{(m)}(k,l) \Bigg[\frac{B_{ij}^{(m)}}{\alpha_{ij}^{(m)}} \sin\big(\omega \alpha_{ij}^{(m)}\big) + \\ c^{-2} \Bigg(\frac{2\omega}{(\alpha_{ij}^{(m)})^2} \cos\big(\omega \alpha_{ij}^{(m)}\big) + \Bigg(\frac{\omega^2}{\alpha_{ij}^{(m)}} - \frac{2}{(\alpha_{ij}^{(m)})^3} \Bigg) \sin\big(\omega \alpha_{ij}^{(m)}\big) \Bigg) - \\ c^{-1} C_{ij}^{(m)} \Bigg(\frac{\sin\big(\omega \alpha_{ij}^{(m)}\big)}{(\alpha_{ij}^{(m)})^2} - \frac{\omega \cos\big(\omega \alpha_{ij}^{(m)}\big)}{\alpha_{ij}^{(m)}} \Bigg) \Bigg]_{\omega_c - B/2}^{\omega_c + B/2}.\end{aligned} \tag{D.35}$$

The CRLB is the given by

$$CRLB(\mathbf{x}) = J(\mathbf{x})^{-1}. \tag{D.36}$$

GREEDY ALTERNATIVE FOR ROOM GEOMETRY ESTIMATION FROM ACOUSTIC ECHOES: A SUBSPACE-BASED METHOD

Mario Coutino[1], Martin Bo Møller[2,3], Jesper Kjær Nielsen[2,3], Richard Heusdens[1]
[1]Delft University of Technology, The Netherlands
[2]Bang & Olufsen A/S, Denmark
[3]Aalborg University, Denmark

***Abstract*— In this paper, we present a greedy subspace method for the acoustic echoes labeling problem, which occurs in applications such as source localization and room geometry estimation. The orthogonal projection into the kernel of the microphones position matrix is used to filter and sort all possible combinations of echoes. A greedy strategy, based on the rank constraint of Euclidean distance matrices (EDMs), is used on the sorted subset of combinations of echoes to extract the feasible combinations. Numerical simulations using room impulse responses (RIRs) from shoe-box shaped rooms show that the method provides improvements in terms of computational complexity and number of required measurements with respect to a recently published graph-based method.**

***Index Terms*— acoustic echoes, room geometry, sorting reflections, greedy algorithm, source localization**

I. INTRODUCTION

In the past years there has been an increasing interest in mapping the shape of a room using acoustic echos [1]-[3]. Knowledge of the room shape can benefit a large number of applications. For example, in the creation of personal sound zones [4][5] one needs to know the room impulse response (RIR) in different locations, which could be modeled if the geometry information of the room is available. In autonomous navigation, knowledge of the enclosure boundary aids collision avoidance. For speech enhancement, knowledge of walls reflections is desirable to compensate for reverberation.

Echoes generated by sound reflected from the room walls carry information about the geometry of the enclosure. By modeling room reflections using virtual sources [6], it is possible to exploit the geometry duality of this representation to estimate the room boundaries. For this purpose, several methods have been proposed to estimate the room geometry with high accuracy. Most of these methods assume knowledge of the RIRs. In [7], the shape in the 2D case is estimated by a single RIR. Antonacci et al. [8] solve the 2D problem assuming multiple sources and microphones.

In instances where multiple microphones, randomly placed in the room, are used to detect the acoustic echos in the RIRs, ambiguities arise at the moment of labeling the echoes according to the wall which produces them. This problem is illustrated in Fig. 1. In order to deal with this issue, Dokmanić et al. in [9] exploits the properties of EDMs to find the room geometry in the general 3D case. More recently, a newly proposed method [2] by Jager et al. has been shown to provide the same accuracy as Dokmanić's method at a much lower computational complexity. This approach recasts the labeling problem of the acoustic echoes problem into a graph problem, which can be solved in reasonable time for instances with a small number of microphones. However, both [9] and [2] become intractable for increasing number of microphones.

In this paper, we aim to build on previous work to further improve the current state-of-the-art solution for the acoustic echo labeling problem. We propose a subspace-based filtering to reduce the computational complexity of the graph-based approach of [2]. Furthermore, we devise a greedy strategy which attains comparable performance to the graph-based method at a reduced computational cost. In addition, the proposed method only requires measurements from a single source, in contrast to the current state-of-the-art method that requires more than one source. In this work, we restrict ourselves to shoe-box shaped rooms as they are commonly found in typical audio reproduction scenarios. However, the proposed method can be extended to arbitrary room geometries.

This work was partially supported by CONACYT

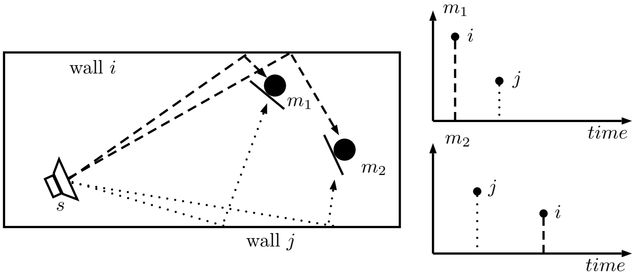


Fig. 1: Ambiguity in the echoes labels due to different order of arrival of wall reflections

II. DATA MODEL

First, let us consider an arbitrary set $\mathcal{M}$ of M microphones located at random positions. That is, $\mathcal{M} = \{\mathbf{r}_m = [x_m, y_m, z_m]^T \in \mathbb{R}^3\}_{m=1}^M$. These locations are assumed known up to a non-singular transformation. Furthermore, consider the set $\mathcal{S} = \{\mathbf{s}_n = [X_n, Y_n, Z_n]^T \in \mathbb{R}^3\}_{n=1}^N$ of N image sources. The squared distances $\mathcal{D} = \{d_{m,n}\} \forall (m,n) \in [1,\dots,M] \times [1,\dots,N]$ between the image sources $\mathcal{S}$ and receivers $\mathcal{M}$ can be measured, i.e., the time-of-arrival (TOA) of the reflections can be estimated at the microphones. Hence, the squared distance $d_{m,n}$ for the (m,n)-th pair can be written as

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

This can be expressed as an inner product as [10]

$$\mathbf{R}_m^T \mathbf{S}_n = d_{m,n} \qquad (2)$$

where the two vectors $\mathbf{R}_m$ and $\mathbf{S}_n$ are given by

$$\mathbf{R}_m = [\mathbf{r}_m^T \mathbf{r}_m \; -2x_m \; -2y_m \; -2z_m \; 1]^T \in \mathbb{R}^{5\times 1}, \qquad (3)$$

$$\mathbf{S}_n = [1 \; X_n \; Y_n \; Z_n \; \mathbf{s}_n^T \mathbf{s}_n]^T \in \mathbb{R}^{5\times 1} \qquad (4)$$

Collecting all the squared distances $d_{m,n}$ for the pairs (m,n) leads to the distance matrix $\mathbf{D} \in \mathbb{R}^{M\times N}$, and the model can be written in matrix form as

$$\mathbf{R}^T \mathbf{S} = \mathbf{D} \in \mathbb{R}^{M\times N} \qquad (5)$$

where $\mathbf{R} = [\mathbf{R}_1, \ldots, \mathbf{R}_M]$ and $\mathbf{S} = [\mathbf{S}_1, \ldots, \mathbf{S}_N]$ are known the microphones and unknown image sources positions matrices, respectively. Even when the positions of the microphones are known up to an arbitrary non-singular matrix $\mathbf{H} \in \mathbb{R}^{5\times 5}$, and the transformed microphones positions matrix $\hat{\mathbf{R}}^T = \mathbf{R}^T\mathbf{H}$ is known instead of $\mathbf{R}$, the model in (5) still holds as

$$\hat{\mathbf{R}}^T\mathbf{H}^{-1}\mathbf{H}\hat{\mathbf{S}} = \mathbf{D}. \tag{6}$$

where $\hat{\mathbf{S}} = \mathbf{H}^{-1}\mathbf{S}$ is the transformed matrix of sources positions.

III. Labeling Acoustic Echoes

From the model in (5), the unknown matrix $\mathbf{S}$ with the position of the sources can be estimated by means of least squares given that $\text{rank}(\mathbf{R}) \geq 5$ when the positions of the receivers and the distance matrix $\mathbf{D}$ are known. However, in most cases, the squared distances in $\mathcal{D}$ are not grouped accordingly to the sources that originate them. That is, the subindex n from the elements in $\mathcal{D}$ is unknown to us. Therefore, an approach to generate $\mathbf{D}$ from the unlabeled set $\mathcal{D}$ is needed.

In this work, we consider the projection into the $\ker(\mathbf{R})$, i.e., kernel of $\mathbf{R}$, to filter and sort all possible combinations of echoes. This projection exploits the structure in the model (5) to estimate the matrix $\mathbf{D}$ from the unlabeled data $\mathcal{D}$. The goal of this approach is to deliver a complexity reduction similar to the one achieved in [2] allowing us to deal with larger instances of the problem generated either by a larger number of microphones and sources, or by uncertainties in the set $\mathcal{D}$.

When proper diversity in $\mathbb{R}^3$ is assumed for the microphone positions, i.e., non co-located locations for the receivers, the only constraint needed in the method to ensure the rank-5 property of the distance matrix $\mathbf{D}$ is $M \geq 5$ [10][11].

A. Subspace Filtering

Let the rank-5 economy-sized SVD of the known receivers position matrix $\mathbf{R}$ be given by

$$\mathbf{R} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T. \tag{7}$$

The complementary orthogonal projection $\Pi^\perp_{\mathbf{R}}$ into $\ker(\mathbf{R})$ can then be computed from the SVD in (7) as

$$\Pi^\perp_{\mathbf{R}} = \mathbf{I}_M - \mathbf{V}\mathbf{V}^T \tag{8}$$

This projection can be shown to have the property

$$\Pi^\perp_{\mathbf{R}}\mathbf{R}^T = 0, \tag{9}$$

hence from (5) it follows that

$$\Pi^\perp_{\mathbf{R}}\mathbf{R}^T\mathbf{S} = \Pi^\perp_{\mathbf{R}}\mathbf{D} = \mathbf{0}, \tag{10}$$

for $\mathbf{D}$-matrices with the correct sorting. In this work (10) is used to estimate $\mathbf{D}$ from $\mathcal{D}$. An interesting property of the complementary projection matrix is that

$$||\Pi^\perp_{\mathbf{R}}||_2 = 1 \tag{11}$$

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

$$\begin{aligned} ||\Pi^\perp_{\mathbf{R}}(\mathbf{D}_c + \mathbf{n})||_2 &= ||\Pi^\perp_{\mathbf{R}}(\mathbf{R}^T\mathbf{S}_c + \mathbf{n})||_2 && (12)\\ &= ||\Pi^\perp_{\mathbf{R}}\mathbf{n}||_2 && (13)\\ &\leq ||\mathbf{n}||_2 && (14)\end{aligned}$$

where $\mathbf{D}_c$ is the c-th column of the matrix $\mathbf{D}$. This makes the projection particularly useful in cases where the elements of $\mathcal{D}$ are perturbed with noise.

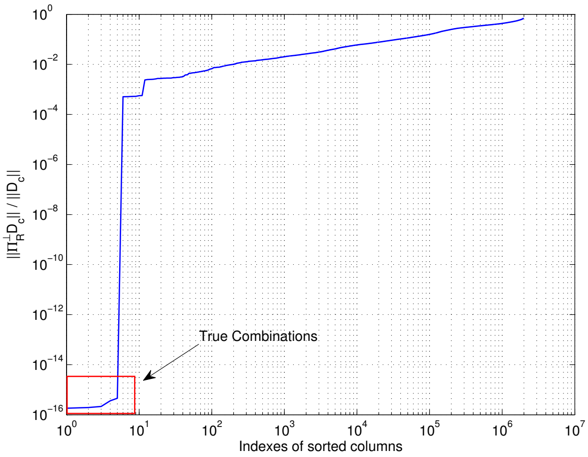


Fig. 2: Normalized functional (16) for the columns $\tilde{\mathbf{D}}$, in the noise free case, sorted in ascending order. In this example $M = 9$ and $N = 5$.

In order to apply the projection given in (8), we first consider the matrix $\tilde{\mathbf{D}}$ defined as the distance matrix generated by all the possible combinations of the elements in $\mathcal{D}$, e.g.,

$$\tilde{\mathbf{D}} = \begin{bmatrix} d_{1,1} & \cdots & d_{1,2} & \cdots & d_{1,N} \\ d_{2,1} & \cdots & d_{2,1} & \cdots & d_{2,N} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ d_{M,1} & \cdots & d_{M,2} & \cdots & d_{M,N} \end{bmatrix} \in \mathbb{R}^{M\times N^M} \tag{15}$$

In the ideal case, i.e., measurements without any kind of noise, the results are straightforward. By defining the functional

$$f(c) = ||\Pi^\perp_{\mathbf{R}}\tilde{\mathbf{D}}_c||_2^2 \;\forall\; c \;\in\; [1, \ldots, N^M] \tag{16}$$

we can select the subset of feasible columns as

$$\mathcal{C} = \{c \,:\, f(c) = 0\}, \tag{17}$$

and provide an estimate of the feasible distance matrix given by

$$\hat{\mathbf{D}} = \tilde{\mathbf{D}}_{\mathcal{C}} \in \mathbb{R}^{M\times N} \tag{18}$$

where $\tilde{\mathbf{D}}_{\mathcal{C}}$ represents the trimmed distance matrix, which only retains the columns specified by the set $\mathcal{C}$. The functional is illustrated in Fig. 2 for a problem instance with $M = 9$ microphones and $N = 5$ (image) sources.

However, in real applications there is no guarantee that the true distances $\mathcal{D}$ are measured perfectly, hence the set in (17) will, most likely, turn out empty. In order to deal with noisy measurements, we provide a column-dependent upper bound for the proposed functional which considers the effect of perturbations.

Consider that the measured squared distance $\hat{d}_{m,n}$ can be expanded as

$$\hat{d}_{m,n} \triangleq (\sqrt{d_{m,n}} + w_{m,n})^2 = d_{m,n} + 2\sqrt{d_{m,n}}w_{m,n} + w_{m,n}^2 \tag{19}$$

where $w_{m,n}$ is the perturbation in the (m, n)-pair measurement. After the projection is applied to a stacked version of (19), the following residual is obtained

$$\Pi^\perp_{\mathbf{R}}\hat{\mathbf{D}}_c = \Pi^\perp_{\mathbf{R}}\left[2\text{diag}(\mathbf{w}_c)\mathbf{D}_c^{\circ\frac{1}{2}} + \text{diag}(\mathbf{w}_c)\mathbf{w}_c\right] \in \mathbb{R}^{M\times 1} \tag{20}$$

where $\mathbf{A}^{\circ p}$ denotes the p-th Hadamard power of the matrix $\mathbf{A}$. Therefore, it is possible to provide a selection rule similar to (17) by upper bounding the square norm of the expression in (20). An appropriate upper bound for the residual norm can be given by

$$\begin{aligned} \|\Pi_{\mathbf{R}}^{\perp}\hat{\mathbf{D}}_c\|_2^2 &= \|\Pi_{\mathbf{R}}^{\perp}[\operatorname{diag}(\mathbf{w}_c)(2\mathbf{D}_c^{\circ\frac{1}{2}} + \mathbf{w}_c)]\|_2^2 && (21)\\ &\leq \|\Pi_{\mathbf{R}}^{\perp}\|_2^2\|\operatorname{diag}(\mathbf{w}_c)\|_2^2\|2\mathbf{D}_c^{\circ\frac{1}{2}} + \mathbf{w}_c\|_2^2 && (22)\\ &\leq 4\max^2(\mathbf{w}_c)\|\mathbf{D}_c^{\circ\frac{1}{2}} + 0.5\mathbf{w}_c\|_2^2 && (23)\\ &= \kappa_c && (24) \end{aligned}$$

where $\max(\mathbf{a})$ denotes the maximum absolute value of the vector $\mathbf{a}$, and the fact that $\|\Pi_{\mathbf{R}}^{\perp}\| = 1$ has been used. Using the derived upper bound, we can build the subset of columns for the distance matrix as

$$\mathcal{C} = \{c : f(c) \leq \kappa_c\} \qquad (25)$$

and estimate the distance matrix using expression (18). Although the bound always holds, κ_c is not directly available from the measurements. As in practice, we deal with realizations of the measurement process, in order to use the bound in (24) we introduce

$$\kappa_c^{(i)} = 4\gamma^i\sigma_{\mathbf{w}}^2\|\hat{\mathbf{D}}_c^{\circ\frac{1}{2}}\|_2^2, \;\; \gamma \geq 1 \qquad (26)$$

as surrogate to provide a practical iterative threshold for the functional. In this expression, $\sigma_{\mathbf{w}}^2$ denotes the noise power. The power of the noise can be assumed known as it is considered that the accuracy of the method employed for obtaining the TOA estimates is known. For simplicity, we consider that all columns are subject to the same noise level σ_w. This assumption affects the performance of the bound as sources located at different positions have different accuracy levels. However, this can be considered a reasonable assumption as the ordering of echoes is unknown. As in simulations it has been observed that $\kappa_c^{(0)}$ is sufficient for the method to deliver adequate results, our algorithm fixes κ_c^i to κ_c^0.

B. Avoiding the Graph Problem

For real measurements, $|\mathcal{C}| \gg N$. Therefore, further processing is required to only select feasible columns. For this step two possible strategies can be applied: (i) the recently proposed graph-based method from [2], where the problem is recast as a maximum independent set problem, or (ii) a greedy approach that sequentially selects feasible combinations. In the following, we solely focus on (ii) as we want to avoid solving the NP-hard problem of listing all maximal independents sets.

To avoid the graph-problem, firstly we make the observation that when using the functional $f(c)$ for sorting the columns of $\tilde{\mathbf{D}}$, the columns of the lowest normalized functional value, meeting the rank constraint for EDMs, most likely belong to the true distance matrix. For this, consider the matrix $\mathbf{E} \in \mathbb{R}^{M\times M}$ as the EDM constructed from the relative distances between receivers. The matrix $\tilde{\mathbf{E}}_c \in \mathbb{R}^{M\times(M+1)}$ denotes an augmented EDM built by adding the column vector $\hat{\mathbf{D}}_c$. The rank of $\tilde{\mathbf{E}}$ is larger than five, if $\mathbf{E}$ is augmented with distances to different sources. As suggested in [2], the ϵ-rank defined as [12]

$$\operatorname{rank}(\tilde{\mathbf{E}}, \epsilon) = \min_{\|\tilde{\mathbf{E}}-\mathbf{X}\|_2\leq\epsilon} \operatorname{rank}(\mathbf{X}) \qquad (27)$$

can be employed to sequentially exclude echoes combinations that, approximately, violate the rank constraint. As the threshold ϵ is unknown a priori, an iterative approach is employed to obtain the suitable candidate for ϵ.

Secondly, as pointed out in [2], the columns in $\hat{\mathbf{D}}$ have unique elements, so in addition to the sequential exclusion of columns by the ϵ-rank constraint, columns sharing elements with already selected feasible columns are rejected. The sub-optimal algorithm combining these two observations is presented in Algorithm 1. The $\eta > 1$ parameter controls the growth of the rank constraint. This allows the solution to only include the best ranked columns.

Algorithm 1 Subspace-based Greedy Acoustic Echoes Sorting

```
Input: $\mathcal{D}$, $\Pi_{\mathbf{R}}^{\perp}$, $\mathbf{E}$, $\epsilon_0$, $N$, $\sigma_{\mathbf{w}}$
Output: $\mathbf{D}$
  Initialization: Generate $\tilde{\mathbf{D}}$ and $\boldsymbol{\kappa}^{(0)}$, $\mathbf{D} = \{\}$, $\epsilon = \epsilon_0$
  $\mathcal{C} = \{c : f(c) \leq \boldsymbol{\kappa}_c^{(0)}\}$
  $\mathcal{C}_s = \text{sort}(\mathcal{C}, f(c)/\|\tilde{\mathbf{D}}_c\|_2^2, \text{"ascending"})$
  $\hat{\mathbf{D}} = \tilde{\mathbf{D}}_{\mathcal{C}_s}$
  while numCols($\mathbf{D}$) $< N$ do
      for $c = 1$ to $|\mathcal{C}_s|$ do
          $\tilde{\mathbf{E}} = \begin{bmatrix} \mathbf{E} & \hat{\mathbf{D}}_c \\ \hat{\mathbf{D}}_c^T & 0 \end{bmatrix}$
          if rank($\tilde{\mathbf{E}}, \epsilon$) $\leq 5$ and $\hat{\mathbf{D}}_c \cap \mathbf{D} == \emptyset$ then
              $\mathbf{D} = [\mathbf{D}, \hat{\mathbf{D}}_c]$
          end if
      end for
      if numCols($\mathbf{D}$) $< N$ then
          $\epsilon = \eta\epsilon$
      end if
  end while
```

Finally, after the matrix $\mathbf{D}$ is estimated by the greedy approach, the least squares solution for the estimates of the source locations, for $M \geq 5$, can be directly obtained by

$$\hat{\mathbf{S}} = (\mathbf{R}^T)^{\dagger}\mathbf{D}. \qquad (28)$$

Contrary to (i), where more than one maximum independent set can be found in the graph, (ii) provides a unique solution. The unique solution allows the echoes to be sorted even when the constraint imposed by Polleyfey's method [10] used in [2] is not met.

Notice that if measurements from Q acoustic sources are available, i.e.,

$$\mathbf{D}_{\text{Tot}} = \left[\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_Q\right] = \mathbf{R}^T\left[\mathbf{S}_1, \ldots, \mathbf{S}_Q\right]. \qquad (29)$$

A combination of Polleyfey's method, using the SVD of $\mathbf{D}_{\text{Tot}}$, and Procrustes analysis can be performed to estimate the image source positions instead of using (28). This approach could lead to better reconstruction results for cases in which the pseudo-inverse of $\mathbf{R}^T$ is not well conditioned.

IV. NUMERICAL SIMULATIONS

In this section results from numerical simulations comparing the proposed greedy method and a modified version of [2] are presented. First, to evaluate the proposed method we generated a set of 500 Monte Carlo simulations for different uncertainties in the measured distances. The simulation illustrates the acoustic echo labeling problem from multiple room reflections, i.e., $N = 6$ for a 3D shoe-box shaped room. The number of microphones considered is $M = 9$. The noise-free distances from the reflections of the walls are taken from the peaks in the simulated impulse responses (RIRs) generated by the acoustics simulation software [13]. As the graph-based method requires multiple sources to provide an estimate of the source positions, a version with an oracle is used instead, i.e., if more than one maximal independent set is found, the closest set (in the least square sense) with respect to the noisy distance matrix is considered as the solution of the method. To provide a speed up to the method, the subspace filtering step is added in this modified

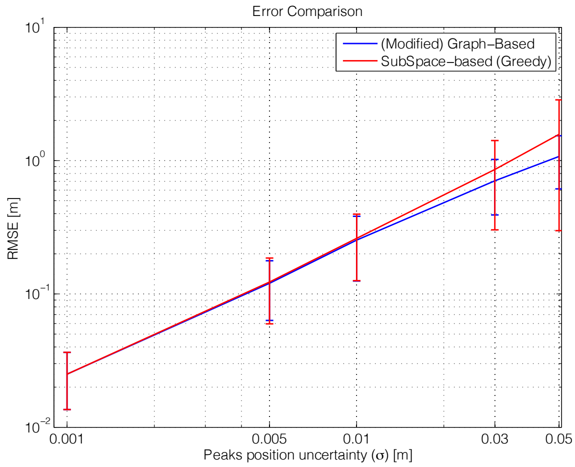


Fig. 3: Estimation error comparison for $M = 9$ and $N = 6$. Error bars show results within one standard deviation.

version. The addition of the subspace filtering to the method shows that it is possible to deliver a feasible implementation of the graph-based approach for relatively large number of microphones, contrary to the intractability statement given in [14].

In Fig. 3 the estimation error of both methods is compared. The error is computed as the norm of the Euclidean distance between the true $\mathbf{s}_n$ and estimated position $\hat{\mathbf{s}}_n$ of the sources, i.e., $\|\mathbf{e}\|_2 = \| [\mathrm{dist}_E(\mathbf{s}_1, \hat{\mathbf{s}}_1), \ldots, \mathrm{dist}_E(\mathbf{s}_N, \hat{\mathbf{s}}_N)]^T \|_2$, where the estimated positions are found by (28) assuming $\mathbf{R}$ known (up to a non-singular transform). Notice how the accuracy of the estimation decreases as the uncertainty in the distances increases. For low uncertainties, i.e., $\sigma < 0.01$m, their accuracy is identical. However, as the uncertainty increases, the results in Fig. 3 show higher degradation of the greedy approach due to its sub-optimality.

A comparison of the relative running time of each method with respect the baseline case of $M = 5$ using the graph-based approach is shown in Fig. 4. For this comparison 500 Monte Carlo simulations were made using different number of microphones. The time consumed by the methods with subspace filtering is considerable lower than the graph-based approach. This result shows that it is possible to find tractable solutions for larger instances of the graph problem by adding subspace filtering. By prefiltering the combinations using the proposed functional, the number of computed SVDs reduces drastically. In addition, by removing combinations that might not be rejected by the rank constraint, the number of nodes in the graph, used to find the maximum independent sets, is reduced. Hence, the method gains an additional speed up. The reduction in time when the number of microphones increases from $M = 5$ to $M = 6$ for the methods with subspace filtering is explained by the selectivity of the kernel of $\mathbf{R}$. As more combinations of echoes are rejected by the subspace filtering, less ϵ-rank checks are performed to obtain a feasible distance matrix $\mathbf{D}$.

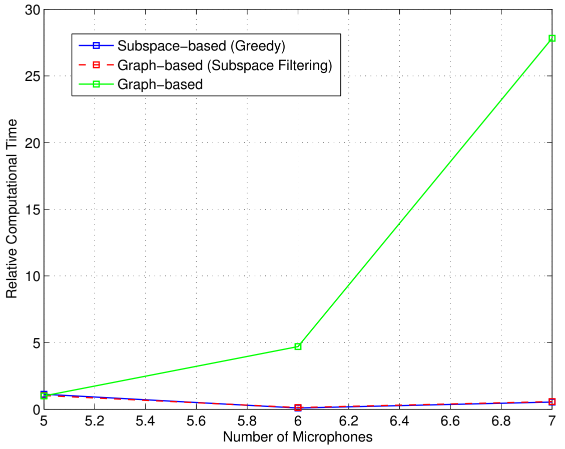


Fig. 4: Comparison of computation time between the graph-based methods and greedy approach for number of microphones.

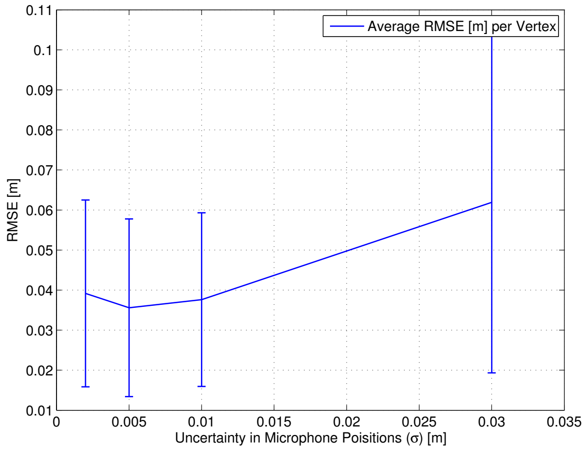


Fig. 5: Average RMSE per vertex of the room for the proposed greedy strategy for different uncertainties in the locations of the microphones. Error bars show results within one standard deviation.

Finally, Fig. 5 illustrates the performance of the proposed method for different uncertainties in the locations of the microphones. For this experiment, 500 simulated measurements were produced from four different acoustic sources. The reconstruction of the image sources positions is done using the noisy locations of the microphones and Pollefey's method [10]. In this experiment, it is assumed that distances between each microphone-image source pair contain additive white Gaussian noise with standard deviation $\sigma_{RIR} = 1$cm. Notice how even in the presence of noise, in both RIRs peaks and microphones locations, the method provides vertex estimates with average RMSE close to 5cm. The high dependency on the accuracy of the positions of the microphones is seen in the increased standard deviation of the RMSE and its mean value.

All numerical simulations were run on a Macbook Air (Mid 2013) 1.7 GHz Inter Core i7 using non-optimized MATLAB code.

V. CONCLUSIONS

In this paper we proposed an alternative approach for the acoustic echoes labeling problem. Using a complementary orthogonal projection related with the receivers locations, it is possible to elucidate a filtering and sorting criteria for the columns of the distance matrix built from all possible combinations of available echoes. It is shown, that in the noise free case perfect identification of the true columns can be achieved. Furthermore, for the noisy case, a greedy alternative is proposed to avoid the solution of the NP-hard problem of listing all maximal independent sets in a graph. Numerical simulations show the applicability of the method and the benefits of applying the subspace filtering to the original graph-based method. In addition, effects of uncertainties in the distance measurements, not discussed in literature before, were shown through numerical experiments.

REFERENCES

[1] Kreković, Miranda, Ivan Dokmanić, and Martin Vetterli. ”EchoSLAM: Simultaneous localization and mapping with acoustic echoes.” 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016.

[2] Jager, Ingmar, Richard Heusdens, and Nikolay D. Gaubitch. ”Room geometry estimation from acoustic echoes using graph-based echo labeling.” 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016.

[3] Dokmanić, Ivan, Laurent Daudet, and Martin Vetterli. ”From acoustic room reconstruction to slam.” 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016.

[4] Betlehem, Terence, et al. ”Personal Sound Zones: Delivering interface-free audio to multiple listeners.” IEEE Signal Processing Magazine 32.2 (2015): 81-91.

[5] Jacobsen, Finn, et al. ”A comparison of two strategies for generating sound zones in a room.” 18th International Congress on Sound and Vibration. 2011.

[6] Allen, Jont B., and David A. Berkley. ”Image method for efficiently simulating small room acoustics.” The Journal of the Acoustical Society of America 65.4 (1979): 943-950.

[7] Dokmanić, Ivan, 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). IEEE, 2011.

[8] Antonacci, Fabio, et al. ”Inference of room geometry from acoustic impulse responses.” IEEE Transactions on Audio, Speech, and Language Processing 20.10 (2012): 2683-2695.

[9] Dokmanić, Ivan, et al. ”Acoustic echoes reveal room shape.” Proceedings of the National Academy of Sciences 110.30 (2013): 12186-12191.

[10] Marc Pollefeys, and David Nister. ”Direct computation of sound and microphone locations from time-difference-of-arrival data.” ICASSP, pp. 2445-2448. 2008.

[11] Gower, John Clifford. ”Properties of Euclidean and non-Euclidean distance matrices.” Linear Algebra and its Applications 67 (1985): 81-97.

[12] G.H. Golub and C.F. Van Loan, Matrix Computations, North Oxford Academic, Oxford, third edition, 1983

[13] E. A. Habets, Room impulse response generator, Technische Universiteit Eindhoven, Tech. Rep 2, 1 (2006).

[14] Ingmar Jager. ”Room Shape Estimation from Acoustic Echoes using Graph-based Echo Labeling.” MSc. Thesis., TU Delft, Delft University of Technology, 2015.

Bibliography

[1] 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, 1979.

[2] Fabio Antonacci, Jason Filos, Mark RP Thomas, Emanuël AP Habets, Augusto Sarti, Patrick A Naylor, and Stefano Tubaro. Inference of room geometry from acoustic impulse responses. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(10):2683–2695, 2012.

[3] Niccoló Antonello, Toon van Waterschoot, Marc Moonen, and Patrick A Naylor. Source localization and signal reconstruction in a reverberant field using the fdtd method. In *Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European*, pages 301–305. IEEE, 2014.

[4] Jacob Benesty, Jingdong Chen, and Yiteng Huang. *Microphone array signal processing*, volume 1. Springer Science & Business Media, 2008.

[5] Terence Betlehem, Wen Zhang, Mark A Poletti, and Thushara D Abhayapala. Personal sound zones: Delivering interface-free audio to multiple listeners. *IEEE Signal Processing Magazine*, 32(2):81–91, 2015.

[6] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[7] Marco Crocco, Alessio Del Bue, and Vittorio Murino. A bilinear approach to the position self-calibration of multiple sensors. *Signal Processing, IEEE Transactions on*, 60(2):660–673, 2012.

[8] Giovanni Del Galdo, Oliver Thiergart, Tobias Weller, and Emanuël AP Habets. Generating virtual microphone signals using geometrical information gathered by distributed arrays. In *Hands-free Speech Communication and Microphone Arrays (HSCMA), 2011 Joint Workshop on*, pages 185–190. IEEE, 2011.

[9] Reinhard Diestel. Graph theory, vol. 173 of. *Graduate Texts in Mathematics*, 2005.

[10] Ivan Dokmanić, Yue M Lu, and Martin Vetterli. Can one hear the shape of a room: The 2-d polygonal case. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 321–324. IEEE, 2011.

[11] Ivan Dokmanic, Reza Parhizkar, Juri Ranieri, and Martin Vetterli. Euclidean distance matrices: Essential theory, algorithms, and applications. *Signal Processing Magazine, IEEE*, 32(6):12–30, 2015.

[12] Ivan Dokmanić, Reza Parhizkar, Andreas Walther, Yue M Lu, and Martin Vetterli. Acoustic echoes reveal room shape. *Proceedings of the National Academy of Sciences*, 110(30):12186–12191, 2013.

[13] Robert P Dougherty and Robert W Stoker. Sidelobe suppression for phased array aeroacoustic measurements. *AIAA paper*, 2242:1998, 1998.

[14] Yaron Doweck, Alon Amar, and Israel Cohen. Joint model order selection and parameter estimation of chirps with harmonic components. *Signal Processing, IEEE Transactions on*, 63(7):1765–1778, 2015.

[15] P-A Gauthier, C Camier, Y Pasco, A Berry, E Chambatte, R Lapointe, and M-A Delalay. Beamforming regularization matrix and inverse problems applied to sound field measurement and extrapolation using microphone array. *Journal of Sound and Vibration*, 330(24):5852–5877, 2011.

[16] Gene H Golub and Charles F Van Loan. *Matrix computations*, volume 3. JHU Press, 2012.

[17] John C Gower and Garmt B Dijksterhuis. *Procrustes problems*, volume 3. Oxford University Press Oxford, 2004.

[18] John Clifford Gower. Euclidean distance geometry. *Mathematical Scientist*, 7(1):1–14, 1982.

[19] Richard A Gramann and James W Mocio. Aeroacoustic measurements in wind tunnels using adaptive beamforming methods. *The Journal of the Acoustical Society of America*, 97(6):3694–3701, 1995.

[20] Emanuel AP Habets. Room impulse response generator. *Technische Universiteit Eindhoven, Tech. Rep*, 2(2.4):1, 2006.

[21] Klaus Hartung, Jonas Braasch, and Susanne J Sterbing. Comparison of different methods for the interpolation of head-related transfer functions. In *Audio Engineering Society Conference: 16th International Conference: Spatial Sound Reproduction*. Audio Engineering Society, 1999.

[22] JA Högbom. Aperture synthesis with a non-regular distribution of interferometer baselines. *Astronomy and Astrophysics Supplement Series*, 15:417, 1974.

[23] Finn Jacobsen, Martin Olsen, Martin Møller, and Finn T Agerkvist. A comparison of two strategies for generating sound zones in a room. In *18th International Congress on Sound and Vibration*, 2011.

[24] Ingmar Jager. *Room Shape Estimation from Acoustic Echoes using Graph-based Echo Labeling*. PhD thesis, TU Delft, Delft University of Technology, 2015.

[25] Ingmar Jager, Richard Heusdens, and Nikolay D Gaubitch. Room geometry estimation from acoustic echoes using graph-based echo labeling.

[26] Jesper Rindom Jensen, Jesper Kjær Nielsen, Mads Graesboll Christensen, and Soren Holdt Jensen. On frequency domain models for tdoa estimation. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 11–15. IEEE, 2015.

[27] Steven M Kay. Fundamentals of statistical signal processing, volume i: estimation theory. 1993.

[28] Charles H Knapp and G Clifford Carter. The generalized correlation method for estimation of time delay. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 24(4):320–327, 1976.

[29] Heinrich Kuttruff. *Room acoustics*. CRC Press, 2009.

[30] Jeffrey C Lagarias, James A Reeds, Margaret H Wright, and Paul E Wright. Convergence properties of the nelder–mead simplex method in low dimensions. *SIAM Journal on optimization*, 9(1):112–147, 1998.

[31] W Marshall Leach. *Introduction to electroacoustics and audio amplifier design*. Kendall/Hunt Publishing Company, 2003.

[32] Jian Li and Petre Stoica. Angle and waveform estimation via relax. *Aerospace and Electronic Systems, IEEE Transactions on*, 33(3):1077–1087, 1997.

[33] Jim Li and Petre Stoica. Efficient mixed-spectrum estimation with applications to target feature extraction. *Signal Processing, IEEE Transactions on*, 44(2):281–295, 1996.

[34] Stéphane G Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *Signal Processing, IEEE Transactions on*, 41(12):3397–3415, 1993.

[35] KA Marsh and JM Richardson. The objective function implicit in the clean algorithm. *Astronomy and Astrophysics*, 182:174–178, 1987.

[36] Jorge Martinez, Nikolay Gaubitch, and W Bastiaan Kleijn. A robust region-based near-field beamformer. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 2494–2498. IEEE, 2015.

[37] Danielle Moreau, Ben Cazzolato, Anthony Zander, and Cornelis Petersen. A review of virtual sensing algorithms for active noise control. *Algorithms*, 1(2):69–99, 2008.

[38] John N Mourjopoulos. Digital equalization of room acoustics. *Journal of the Audio Engineering Society*, 42(11):884–900, 1994.

[39] Jesper Kjær Nielsen, Tobias Lindstrøm Jensen, Jesper Rindom Jensen, Mads Græsbøll Christensen, and Søren Holdt Jensen. Grid size selection for nonlinear least-squares optimization in spectral estimation and array processing. In *European Signal Processing Conference (eusipco)*, 2016.

[40] Jesper Kjcer Nielsen, Nikolay D Gaubitch, Richard Heusdens, Jorge Martinez, Tobias Lindstrom Jensen, and Soren Holdt Jensen. Real-time loudspeaker distance estimation with stereo audio. In *Signal Processing Conference (EUSIPCO), 2015 23rd European*, pages 250–254. IEEE, 2015.

[41] Marc Pollefeys and David Nister. Direct computation of sound and microphone locations from time-difference-of-arrival data. In *ICASSP*, pages 2445–2448, 2008.

[42] Vikas C Raykar and Ramani Duraiswami. Automatic position calibration of multiple microphones. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, volume 4, pages iv–69. IEEE, 2004.

[43] Bahaa Saleh. *Introduction to subsurface imaging.* Cambridge University Press, 2011.

[44] Amit Shinde, Anshuman Sahu, Daniel Apley, and George Runger. Preimages for variation patterns from kernel pca and bagging. *IIE Transactions*, 46(5):429–456, 2014.

[45] Guy-Bart Stan, Jean-Jacques Embrechts, and Dominique Archambeau. Comparison of different impulse response measurement techniques. *Journal of the Audio Engineering Society*, 50(4):249–262, 2002.

[46] Jenho Tsao and Bernard D Steinberg. Reduction of sidelobe and speckle artifacts in microwave imaging: the clean technique. *Antennas and Propagation, IEEE Transactions on*, 36(4):543–556, 1988.

[47] Giacomo Vairetti, Enzo De Sena, Toon van Waterschoot, Marc Moonen, Michael Catrysse, Neofytos Kaplanis, and Søren Holdt Jensen. A physically motivated parametric model for compact representation of room impulse responses based on orthonormal basis functions. *Proc. of the 10th Eur. Congr. and Expo. on Noise Control Eng.(EURONOISE 2015), Maastricht, The Netherlands*, pages 149–154, 2015.

[48] Giacomo Vairetti, Toon van Waterschoot, Marc Moonen, Michael Catrysse, and Søren Holdt Jensen. Sparse linear parametric modeling of room acoustics with orthonormal basis functions. In *2014 22nd European Signal Processing Conference (EUSIPCO)*, pages 1–5. IEEE, 2014.

[49] Alle-Jan van der Veen and Stefan J Wijnholds. Signal processing tools for radio astronomy. In *Handbook of Signal Processing Systems*, pages 421–463. Springer, 2013.

[50] Toon Van Waterschoot and Geert Leus. Static field estimation using a wireless sensor network based on the finite element method. In *Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2011 4th IEEE International Workshop on*, pages 369–372. IEEE, 2011.

[51] Toon Van Waterschoot and Geert Leus. Distributed estimation of static fields in wireless sensor networks using the finite element method. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 2853–2856. IEEE, 2012.

[52] Yanwei Wang, Jian Li, Petre Stoica, Mark Sheplak, and Toshikazu Nishida. Wideband relax and wideband clean for aeroacoustic imaging. *The Journal of the Acoustical Society of America*, 115(2):757–767, 2004.

[53] Earl G Williams. *Fourier acoustics: sound radiation and nearfield acoustical holography.* Academic press, 1999.

[54] Ilan Ziskind and Mati Wax. Maximum likelihood localization of multiple sources by alternating projection. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 36(10):1553–1560, 1988.