System Parameter Estimation of Acoustic Scenes using First Order Microphones

Master's Thesis

Thomas Wilding

Graz, November 15, 2016

Institute of Electronic Music and Acoustics

University of Music and Performing Arts Graz

Graz University of Technology

Advisor: DI Christian Schörkhuber

Assessor: O.Univ.Prof. Mag.art. DI Dr.techn. Robert Höldrich

institut für elektronische musik und akustik

Abstract

Das Ziel dieser Arbeit ist die Entwicklung von Algorithmen zur akustischen Selbstkalibrierung eines Systems von verteilten, synchronisierten Mikrofonarrays, sowie zur Analyse der Geometrie des umgebenden Raumes. Bei den verwendeten Mikrofonarrays handelt es sich um Mikrofone erster Ordnung (B-Format Arrays bestehend aus vier diskreten Kapseln), mit deren Hilfe das Schallfeld an der Position des jeweiligen Arrays in drei Dimensionen analysiert werden kann.

Im Zuge der Selbstkalibrierung werden die Positionen der Mikrofone sowie die zur Kalibrierung verwendeten akustischen Quellen aus den Aufnahmen der Mikrofonarrays bestimmt. Basierend auf den geschätzten Positionen von Mikrofonen und Kalibrationsquellen wird weiterführend die Raumgeometrie untersucht, wobei diese Arbeit Methoden für sowohl schuhschachtelförmige Räume als auch beliebige Umgebungsgeometrien untersucht.

Um eine einfache Anwendbarkeit zu ermöglichen, werden als Quellsignal einfach wiederhol- und manuell erzeugbare, impulshafte Schallereignisse verwendet, beispielsweise Klatschen. Diese Art der Kalibrationsquellen soll die Ableitung von Pseudo-Impulsantworten aus den Mikrofonsignalen ermöglichen, ohne die oftmals aufwendigen Methoden zur Messung von Impulsantworten. Aus diesen Aufnahmen werden Ankunftszeiten sowie Richtungen von Direktschall und möglichen Reflexionen bestimmt um Algorithmen zu ermöglichen, die sich beide Informationen zu Nutze machen.

In dieser Arbeit werden ausschließlich reale Aufnahmen aus echten Räumen verwendet, alle Berechnungen werden in Matlab durchgeführt.

Abstract

The aim of this work is to develop a toolbox containing algorithms aiming at performing the self-calibration of a system of distributed synchronized microphone arrays and for examining the geometric features of the surrounding acoustic scene (room inference). The microphone arrays to be used are first order microphones (B-format arrays consisting of four discrete capsules) allowing three-dimensional analysis of the surrounding sound field.

The aim of the self-calibration is the retrieval of the microphone positions, with the additional advantage of acquiring the positions of the calibration sources as well. Based on the found microphone and calibration source positions the geometry of the surrounding acoustic scene will be determined, with the examined geometries being either shoe-box shaped rooms or rooms of arbitrary shapes with an arbitrary number of reflective surfaces.

The source signals used allow simple repetition while exhibiting an impulsive character to ensure easy applicability (claps). By the use of these impulsive source signals, pseudo-room impulse responses can be derived from the microphone recordings without the usually rather time consuming procedures needed for impulse response measurements. From the microphone recordings the directions as well as times of arrival of the direct sound and reflections will be estimated, all of which are then used for the self-calibration and room inference tasks.

With the aim of a fast applicability, only real measurements performed in real rooms are used in this work. All computations are performed in Matlab.

Acknowledgments

In the beginning I would like to thank Christian Schörkhuber and Dr. Robert Höldrich for their support and input to this work.

I also want to show my gratitude towards everybody that offered advice and time.

Furthermore I would like to thank the WKO Steiermark for the received scholarship.

I'm endlessly thankful towards my family and girlfriend.

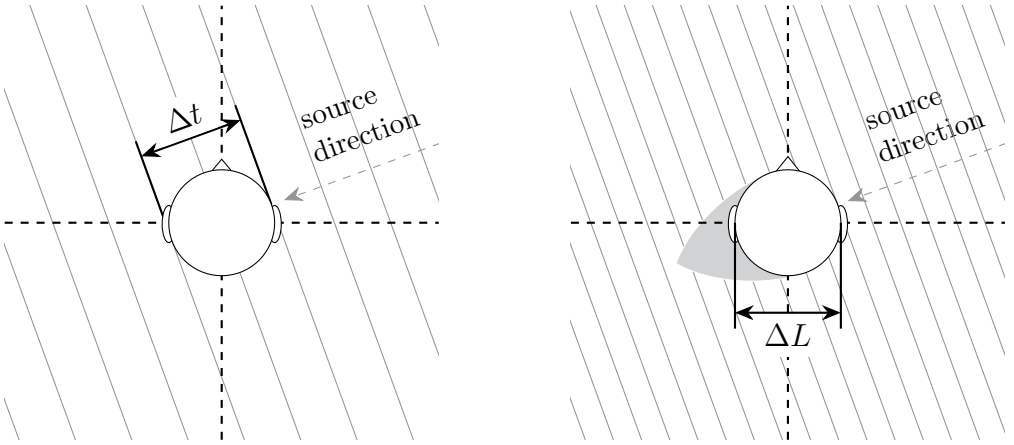
Contents

1 Introduction

In recent years an increasing interest in virtual acoustics also fuelled the research in the field of acoustic room inference and self-calibration of distributed microphones or microphone arrays. Acoustic room inference (or acoustic geometry inference) deals, as the name indicates, with estimating the geometric boundaries of a room or of reflective surfaces in general. To perform this task in a meaningful way, the positions of the microphones need to be known, which in turn usually requires examining the room either with a tape measure or a laser distance measure to determine the microphone positions or at least all inter-microphone distances.

To eliminate these tedious measurements, performing self-calibration of the microphones using only the microphone recordings of an acoustic calibration source located somewhere (in the same room) can be a powerful tool. What self-calibration furthermore presents us with are the positions of the sources that were used for calibration, effectively enabling us to find the locations of points of interest by placing a calibration source there.

The way such algorithms work is usually very similar to what we do when we try to localize an acoustic source (apart from the intuitive part of visual localization). Both our ears pick up the sound of the source, though our left and right ear usually never 'hear' the perfectly same thing, depending on the direction from which the sound of the source arrived. By letting our ears work together we can find inter-aural cues allowing us to guess the correct direction of an acoustic source. Said cues are the *inter-aural time difference* (ITD) and the *inter-aural level difference* (ILD), both of which we use to localize sound sources, though in different frequency ranges, as described in the book by Weinzierl in the chapter on *Spatial Hearing* by Blauert and Braasch [Wei08, Chapter 3]. Due to the size of our head, sound waves below $1600Hz$ (i.e. with wavelengths larger than $\lambda = \frac{c}{f} = \frac{340}{1600} \approx 21\ cm$) can bend around our head, effectively ignoring it with respect to level differences, resulting in time difference based sound source localization. For frequencies above $1600Hz$, the wavelengths become smaller than the distance between our ears, resulting in unpredictable phase errors between the signals arriving at each of our ears, thus prohibiting time based localization. Luckily, shadowing caused by the size of our head reduces the level at the ear averted to the sound source and allows a level based localization in this high frequency band. A visualization of the ITD and ILD can be seen in Figure 1.



(a) ITD **(b)** ILD

Figure 1: Inter-aural time (ITD) and level difference (ILD) in the corresponding frequency ranges below and above $1600Hz$. Δt indicates a time and ΔL a level difference.

The ITD and ILD can be found for microphones in the same way as for our ears, with sound events arriving at distributed microphones at different times and levels (when using directional microphones). The difference between the microphones and our ears is that we have learned our whole life how to localize sound sources, with the additional advantage that we can move our head to increase the accuracy of our localization ability. In the context of array signal processing, the ITD is usually termed *time difference of arrival* (TDOA), which probably is the most important detail observed at a single microphone array, distributed microphones or distributed microphone arrays. To be able to compute the TDOAs, the *time of arrival* (TOA) of sound events at the position of the microphone array have to be determined. A detailed description of the time parameters (TOA and TDOA) is given in Section 1.1. Just like with our ears, the TDOAs at two microphones can be used to find the *direction of arrival* (DOA) of the acoustic signal emitted by an acoustic source. The advantage of microphone arrays distributed within an acoustic scene is the possibility of localizing sources by triangulation from which the distance to the localized source can be determined. Our ears in contrast only allow a vague sense of the distance to a source which we also have to learn for different sources before being able to safely rely on.

Just like with our very own binaural source localization system, we can also perform room inference in a very basic sense. When entering a room we get a good feeling of its size, with a large cathedral sounding much different from our living room. Additionally, we can hear if we are standing or sitting close to a reflective surface,

e.g. a wall, or an absorptive surface of some sort, e.g. a curtain. What we usually cannot hear (at least in an ordinary room) are distinct wall reflections, the reason for which is described by the precedence and Haas effect. The precedence effect states that the direction of the first wave front that arrives at our ears indicates the perceived direction of the sound source and the Haas effect that reflections arriving within the echo threshold, corresponding to a reflection time delay of 1-80 ms, can even be louder (6-10 dB for 20 ms delay) than the direct sound without changing the perceived direction (see [Wei08, Chapter 3] for more details). Still, when we walk through a city we can clearly hear reflections off house walls, especially if the actual source is shielded.

Due to the inherent similarities between our ears and microphones, it should be possible to achieve similar (or in some areas even better) results when examining an acoustic scene using only the signals recorded by distributed microphone arrays, which is the motivation for this work. The main focus lies in the retrieval of information contained in the signals of distributed first order microphones as well as the application in the field of acoustic scene reconstruction. The following section will give a detailed description of the problem dealt with, as well as the surrounding conditions and other assumptions.

1.1 Problem Formulation

This work deals with the *self-calibration* of distributed microphone arrays and with *room inference* (i.e. localizing reflective surfaces) of the room in which the arrays are set up. All this will be combined into a toolbox for analysis of different parameters of an acoustic scene as well as for finding possible ways to use the estimated parameters to create a reconstruction of the corresponding actual scene, with all computations performed in Matlab.

The system parameters tried to estimate in this work are the positions of the microphone arrays and locations and orientation of reflecting surfaces (the room geometry in the wider sense). To find these system parameters, each microphone estimates the TOAs and DOAs of sound events at its spatial location. The TOAs are defined as the points where a sound event (direct sound or reflections) can be found in the recording of a microphone array, e.g. the sample indices, and the DOAs as the directions from which these events arrive at the respective microphone array.

The self-calibration of the distributed microphone arrays (i.e. localizing the microphone arrays and the sources used for calibration) only needs information about the

type of microphone arrays used (e.g. geometry, capsules, ...). Since there do not exist many algorithms that explicitly use TOA and DOA estimates together for the self-calibration task, an algorithm that uses both is proposed. The found TOAs and DOAs, as well as the estimated positions of sources and microphone arrays, are used in the room inference task.

This work can be separated into two main stages and some sub-stages: the *parameter estimation* stage containing the TOA and DOA estimation, and the *scene reconstruction* stage performing the self-calibration and the estimation of reflective surfaces. It will be attempted to keep these stages (as well as the sub-stages) separated, allowing simple adaptions and interchanges of the algorithms.

An acoustic scene model can incorporate a wide range of parameters, some obviously significant (locations of fixed or moving sources, reflector locations, microphone positions, ...) but also others that do not seem as important in the beginning like air temperature ϑ (in $°C$) or wind velocity c_{wind} (in $\frac{m}{s}$) as well as wind direction. All of these can have significant impact on the perception of an acoustic scene of increasing scale. Although there exist approaches for estimating the temperature from microphone recordings (e.g. by Filos in [Fil13]), this will not be considered part of the estimation problem examined in this work.

The acoustic scenes examined in this work will furthermore be restricted to the indoors thus completely removing wind and allowing an assumption of constant room temperature during all the measurements. This simplifications can be combined in form of a *slowly time varying* system allowing the assumption of next to no change of the system during the examination window. Being indoors additionally ensures that there will exist reflections of some sort, and also makes sense for testing and evaluation of an implementation with construction plans available. Alongside the restriction to an indoor scene, the geometries examined in the room inference task can be shoebox-shaped rooms consisting of 4 walls (termed *north*, *south*, *east* and *west*) as well as floor and ceiling, but also rooms with an arbitrary orientation and number of reflecting surfaces.

The microphones used are *Oktava 4D Ambient* microphones (see Appendix A) which are *B-format* microphone arrays consisting of four capsules that output the recorded signals in *A-format* (i.e. as discrete capsule signals). This gives complete control over the following signal processing. Moreover it will be assumed that all the microphones are synchronized, allowing the use of TOA information. Additionally, these microphones were measured extensively by Hack [Hac15] giving a lot of insight into DOA estimation and the resulting source localization capabilities. The *Oktava*

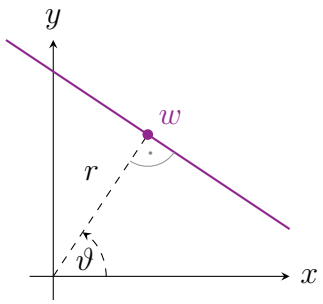


Figure 2: Parameters r and ϑ of a possible reflecting surface

microphones are also used by Schörkhuber et al. [SZZ14] for the WiLMA[1] project that was conducted at the *Institute of Electronic Music and Acoustics* (IEM), allowing straightforward application.

The calibration source signals are assumed to exhibit an impulsive character (i.e. broadband spectrum) with claps being used for all measurements conducted in this work. Furthermore, only a single source is allowed to be active over an examined signal window. The examined signal window should in turn be long enough to include the most interesting part of the room response (i.e. direct sound and first order reflections). The measurements that were carried out are described in detail in Appendix B.

Due to the fact that the orientation of each microphone is assumed to be unknown (in addition to their spatial locations), the DOA of the direct sound of the reference source is used to define the direction of the x-axis in the local coordinate system of the respective microphone. After localization of all sources and microphones computing all points and directions in an arbitrary global coordinate system is trivial.

The reflectors that make up the room geometry are assumed to be linear and are described by the point of the reflector that is closest to the origin in spherical coordinates, shown in Figure 2.

1.1.1 Time and Direction Parameters

The TOAs of the direct sound and the ℓ-th reflection are denoted by $t_{i,j}$ and $t_{i,j}^{(\ell)}$ respectively, representing the time instance at which the sound event of source j can be found in the recording of microphone i. It is important to notice that this work uses two different TDOAs that are derived using these TOAs. The first is the TDOA of the direct sound with respect to different microphone arrays, representing

[1] *wireless large-scale microphone array*

the distance difference the sound emitted by source j has to travel to reach different microphones i. This direct sound TDOA will be denoted by $\Delta t_{i,j}^{(i_0)}$ and can be described by the formula

$$\Delta t_{i,j}^{(i_0)} = t_{i,j} - t_{i_0,j}, \tag{1.1.1}$$

where i_0 is the index of the reference microphone. The reference microphone i_0 of source j is defined as the one closest to that source.

The second definition is the time difference between direct sound and reflections of that direct sound, which is proportional to the additional distance the reflections have to travel to reach the microphone, with the reference in this case being the TOA of the direct sound at the respective array. This reflection TDOA will be denoted by $\Delta \underline{t}_{i,j}^{(\ell)}$ and is computed as

$$\Delta \underline{t}_{i,j}^{(\ell)} = t_{i,j}^{(\ell)} - t_{i,j}. \tag{1.1.2}$$

The difference between the two types of TDOAs is visualized in Figure 3 with the direct sound TDOAs in grey and the reflection TDOAs in the colour of the respective microphone signal. The exemplary signals are shown for four discrete microphones.

Similar to the TDOAs, a *direction difference of arrival* (DDOA) $\Delta \upsilon_{i,j}^{(j_0)}$ of different direct sounds arriving from direction $\upsilon_{i,j}$ is defined by

$$\Delta \upsilon_{i,j}^{(j_0)} = \upsilon_{i,j} - \upsilon_{i,j_0}, \tag{1.1.3}$$

where j_0 indicates the source of which the direct sound direction is used as a reference, i.e. the direction of 0°. A similar DDOA is defined for reflections and the corresponding direct sound, described by

$$\Delta \underline{\upsilon}_{i,j}^{(\ell)} = \upsilon_{i,j}^{(\ell)} - \upsilon_{i,j}, \tag{1.1.4}$$

where $\upsilon_{i,j}^{(\ell)}$ is the DOA of reflection ℓ from source j at microphone i and $\upsilon_{i,j}$ the DOA of the direct sound needed to anchor the unknown orientation of the microphone arrays. All parameters are summarized in Table 1 alongside a description and formula symbol.

The self-calibration will be performed using only the *direct sound* parameters from Table 1. The room inference will then compute reflection points and, building on those, estimate the geometric properties of the room using the respective DDOAs and TDOAs from Table 1 as well as the results from the self-calibration. Both tasks are performed in two dimensions with the extension to three dimensions being straightforward.

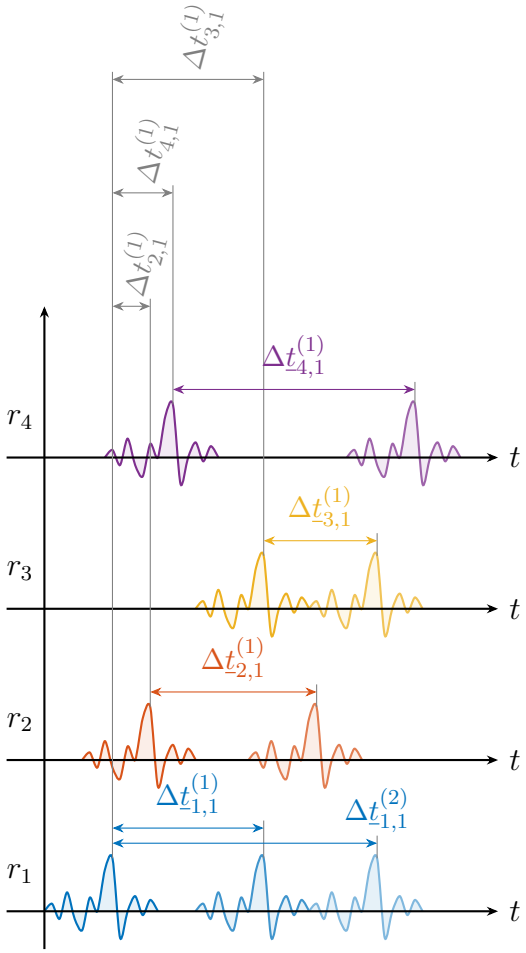


Figure 3: Diagram of the two types of TDOAs, $\Delta t_{i,j}$ (in grey) and $\Delta \underline{t}^{(\ell)}_{i,j}$ (in the respective colours ▮,▮,▮ and ▮), for a source $j = 1$ and its reflections $\ell = \{1, 2\}$ recorded by four different microphones denoted by r_i with $i = \{1, 2, 3, 4\}$ with the reference microphone being the first i.e. $i_0 = 1$. The wave form is an example created by hand assuming ideal reflection.

1.1.2 Notations and Definitions

Lower case bold letters $\mathbf{a}$ are used to represent column vectors while upper case letters $\mathbf{A}$ represent matrices. A signal in the time domain is denoted by $x(n)$, with its frequency domain counterpart $X(k, m)$ obtained by use of the *short-time Fourier transform* (STFT). Signals as column vectors are then again identified with bold letters, i.e. $\boldsymbol{x}(n)$ in the time and $\boldsymbol{X}(k, m)$ in the frequency domain.

The positions of microphones and sources will denoted as vectors $\mathbf{r}_i$ for microphone i and $\mathbf{s}_j$ for source j containing the coordinates in Cartesian or spherical coordinates

Table 1: Parameters and corresponding variables

$t_{i,j}$	TOA of direct sound from source j at microphone i
$t_{i,j}^{(\ell)}$	TOA of reflection ℓ from source j at microphone i
$\Delta t_{i,j}^{(i_0)}$	TDOA of direct sounds at microphone i of source j
$\Delta t_{i,j}^{(\ell)}$	TDOA of reflection ℓ corresponding to source j at microphone i
$\upsilon_{i,j}$	DOA of direct sound j at microphone i
$\upsilon_{i,j}^{(\ell)}$	DOA of reflection ℓ from source j at microphone i
$\Delta\upsilon_{i,j}^{(j_0)}$	DDOA of direct sound j relative to DOA of direct sound j_0
$\Delta\underline{\upsilon}_{i,j}^{(\ell)}$	DDOA of reflection ℓ of source j relative to DOA of direct sound

which can either have two or three dimensions.

$$\mathbf{r}_i = \begin{pmatrix} r_{1,i} & r_{2,i} & r_{3,i} \end{pmatrix}^T \qquad \mathbf{r}_i^\circ = \begin{pmatrix} r_{r,i} & \phi_{r,i} & \theta_{r,i} \end{pmatrix}^T \tag{1.1.5}$$

$$\mathbf{s}_j = \begin{pmatrix} s_{1,j} & s_{2,j} & s_{3,j} \end{pmatrix}^T \qquad \mathbf{s}_j^\circ = \begin{pmatrix} r_{s,j} & \phi_{s,j} & \theta_{s,j} \end{pmatrix}^T \tag{1.1.6}$$

The operators $(\cdot)^T$ and $(\cdot)^H$ represent the transpose and conjugate transpose respectively. Coordinates in two dimensions are also represented by complex numbers (which can be easily expanded to three dimensions using quaternions) resulting in

$$\underline{z} = x + i \cdot y \tag{1.1.7}$$

$$z = |\underline{z}| \cdot e^{i \arg \underline{z}} \tag{1.1.8}$$

$$\mathbf{z} = \begin{pmatrix} \mathrm{Re}\{\underline{z}\} & \mathrm{Im}\{\underline{z}\} \end{pmatrix}^T = \begin{pmatrix} \mathrm{Re}\{z\} & \mathrm{Im}\{z\} \end{pmatrix}^T = \begin{pmatrix} x & y \end{pmatrix}^T \tag{1.1.9}$$

in the complex plane in the general complex form, in polar form or as a vector in Cartesian coordinates respectively.

In the spherical coordinate system used the azimuth and elevation angles will be denoted by ϕ and θ respectively, both starting at the positive x axis with $0°$ and ranging from $-\pi/2$ to $\pi/2$ for ϕ with positive angles towards the $+z$ axis, and from $-\pi$ to π for θ with positive angles towards the $+y$ axis as shown in Figure 4.

The signals recorded by array i and capsule κ with source j active in the examined snapshot, denoted by $x_{i,j}^{(\kappa)}(n)$, are combined in the vector

$$\boldsymbol{x}_{i,j}(n) = \begin{pmatrix} x_{i,j}^{(1)}(n) & x_{i,j}^{(2)}(n) & x_{i,j}^{(3)}(n) & x_{i,j}^{(4)}(n) \end{pmatrix}^T, \tag{1.1.10}$$

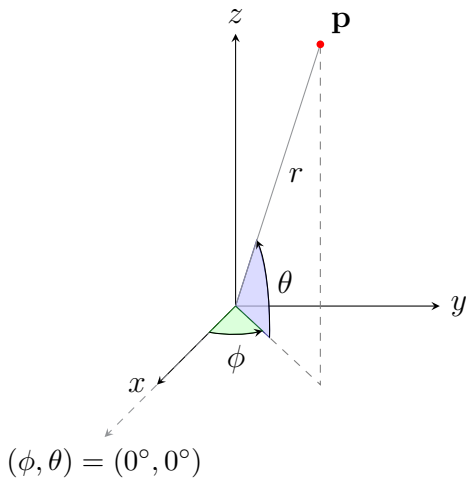


Figure 4: Spherical coordinate system with an examplary point **p**. The angles azimuth ϕ and elevation θ both start with positive angles from the $+x$ axis to the $+y$ and $+z$ axis respectively, negative angles move towards the respective negative axis such that $\phi \in (-\pi, \pi]$ and $\theta \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$.

in the time domain and as

$$\boldsymbol{X}_{i,j}(k,m) = \begin{pmatrix} X_{i,j}^{(1)}(k,m) & X_{i,j}^{(2)}(k,m) & X_{i,j}^{(3)}(k,m) & X_{i,j}^{(4)}(k,m) \end{pmatrix}^T, \tag{1.1.11}$$

in the frequency domain, where $X_{i,j}^{(\kappa)}(k,m)$ is the signal of microphone array i from source j and capsule κ in the frequency domain with k denoting the frequency bin.

1.1.3 Signal Model

The parameters that will be examined are the DOAs and the TOAs (and the resulting TDOAs) of a sound event (direct sound and reflections) at the position of each microphone in a reverberant environment. It is assumed that only a single direct source is active within each examined time interval, i.e. all reflections arriving after the direct sound stem from that very source. In other words, the room is excited by a clap and the next clap will be performed after the room response decayed. The signal model is described in continuous time denoted by the time variable t and only takes into account the direct sound and early reflections (assumed to be *first order reflections*), omitting the late reverberation.

For sound propagation all microphones are assumed to be in the far field of the sources allowing the assumption of a plane wave impinging on each microphone array with respect to the array aperture. For multipath propagation the superposition principle can be applied which leads to the signals at the microphones corresponding

to

$$\tilde{x}_{i,j}(t) = \frac{a_i(\Omega_{i,j})}{r_{i,j}} \cdot s_j(t) + \sum_{\ell=1}^{N_\ell} \frac{a_i(\Omega_{i,j}^{(\ell)})}{r_{i,j}^{(\ell)}} \cdot \delta(t - \Delta\underline{t}_{i,j}^{(\ell)}) * s_j^{(\ell)}(t) \tag{1.1.12}$$

$$= \frac{a_i(\Omega_{i,j})}{r_{i,j}} \cdot s_j(t) + \sum_{\ell=1}^{N_\ell} \frac{a_i(\Omega_{i,j}^{(\ell)})}{r_{i,j}^{(\ell)}} \cdot s_j^{(\ell)}(t - \Delta\underline{t}_{i,j}^{(\ell)}) \tag{1.1.13}$$

with

$$s_j^{(\ell)}(t) = (s_j * w_\ell)(t), \tag{1.1.14}$$

where $s_j(t)$ is the signal of source j, $s_j^{(\ell)}(t)$ is the signal after reflection at reflector ℓ, $w_\ell(t)$ is the corresponding filter of the reflector and $\Delta\underline{t}_{i,j}^{(\ell)}$ is the TDOA of reflection ℓ arriving at microphone i from source j. The operator $*$ denotes the convolution. The entity $r_{i,j}$ describes the radial distance the direct sound has to travel from source j to microphone i and $r_{i,j}^{(\ell)}$ the distance reflection ℓ has to travel. The path the sound travels through air is modelled as a non dispersive ideal channel introducing a simple time delay $h_{air}(t) = \delta(t - \Delta t)$.

The steering factor (describing the array manifold) is included as $a_i(\Omega)$ for microphone i in direction $\Omega = (\phi, \theta)$. The direct sounds arrive from directions $\Omega_{i,j} = (\phi_{i,j}, \theta_{i,j})$ and the reflections from directions $\Omega_{i,j}^{(\ell)} = (\phi_{i,j}^{(\ell)}, \theta_{i,j}^{(\ell)})$ with the usual indices. The final signal recorded by microphone i results in

$$x_i(t) = \tilde{x}_{i,j}(t) + n(t) \tag{1.1.15}$$

with $n(t)$ being the noise model, assumed to be uncorrelated white Gaussian noise, incorporating the microphone as well as background noise.

1.2 Overview

Different parameters that can be found for an acoustic scene are shown in Figure 5, separated into the four groups *microphones*, *sources*, *calibration sources* and *reflectors*, including the parameters inherent to each group. Since the types or models of microphones are either chosen or given, the only parameters needed to estimate are the orientation and the position of each microphone. The reflectors are described by their rotation and the position of a point on the reflector (size and characteristics will not be examined in this work). The parameters that are examined in this work are highlighted in Figure 5. It should be noted that the source positions that are

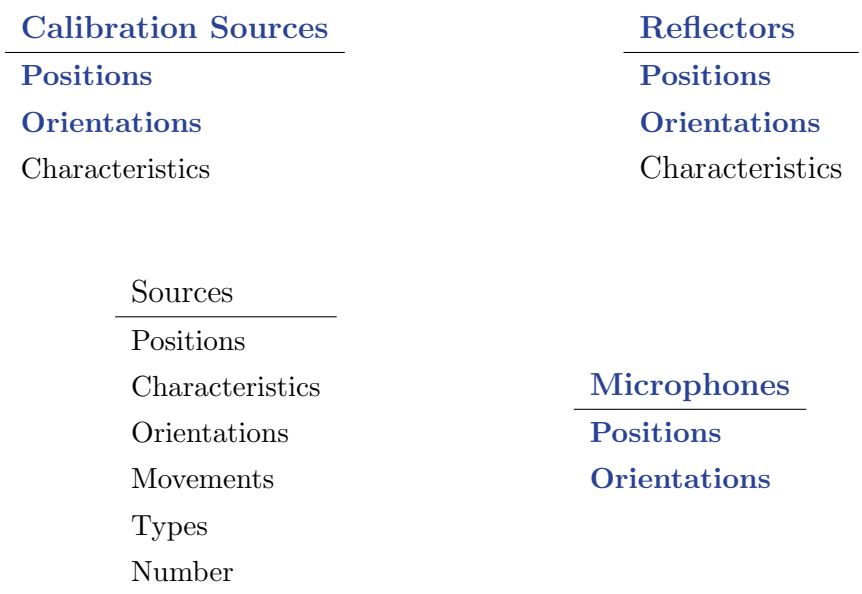


Figure 5: Overview of four groups that can be found in an acoustic scene. The groups and parameters examined by this work are highlighted in **bold blue** letters.

estimated are not those of arbitrary sources, as for example a talking person or a car, but those of calibration sources which should fulfill the restrictions mentioned in Section 1.1.

1.2.1 Algorithm Overview

A flow graph of the algorithm that is described in the following sections can be seen in Figure 6. It contains all important algorithm blocks (described in detail in Chapters 2 and 3) and gives an overview of the whole algorithm as well as illustrating the possibilities to exchange parts of the algorithm. Overall it was intended to keep the algorithms easy to maintain and the different parts as separated as possible. The flow graph also indicates computations that are performed on each microphone array i.

Initially the recorded signals are transformed into the frequency domain using the STFT, where the usual set of options can be chosen (FFT length, window size, hop size and window type). Due to the percussive character of the intended calibration signal (claps) finding a balance between time and frequency resolution is important, with a good time resolution being slightly more important for detecting the TOAs of

direct sound and reflections at the (in the best case) correct time instances.

The *parameter estimation* algorithms (described in Chapter 2) only need the STFT $\boldsymbol{X}_{i,j}(k,m)$ of the signals $\boldsymbol{x}_{i,j}(n)$ recorded by microphone array i resulting in the TOAs of direct sound and reflections as well as the corresponding DOAs. The process of how these estimates are obtained is irrelevant for the following algorithms. The TDOAs and DDOAs at each microphone array are computed sequentially with the TDOAs computed beforehand since they are found with the inclusion of a weighting and are needed for picking the final DOAs at the corresponding TOAs. The TOAs are in the form of functions, termed $\hat{h}_{i,j}(m)$ and $\hat{h}_{\nu,j}(m)$, comparable to room impulse responses (termed *pseudo-room impulse responses* in the following work), with ν indexing all other microphone arrays. To be able to compute time differences, the TOA functions $\hat{h}_{\nu,j}(m)$ of all other microphones ν are needed (i.e. for computing the direct sound TDOAs all direct sound TOAs of all microphones are needed). The estimated DOAs $\hat{\Omega}_{i,j}(m)$ are functions as well, giving a DOA at each frame m. To find the reflection TDOAs and reflection DDOAs no interaction with other microphone arrays is needed, since all these values are referenced to the direct sound at that specific array.

The results of the parameter estimation are the TDOAs of the direct sounds $\Delta t_{i,j}$ and the direct sound DOAs $\upsilon_{i,j}$, as well as the TDOAs of the reflections from the corresponding direct sound $\boldsymbol{\Delta t}_{i,j}^{(\ell)}$ and the reflections DDOAs $\boldsymbol{\Delta \underline{\upsilon}}_{i,j}^{(\ell)}$ to the corresponding direct sound DOA. The latter are denoted by vectors since an arbitrary number of reflections can be detected by each microphone array.

The *scene reconstruction* block (described in Chapter 3) uses all the parameters obtained from each microphone array. It performs the self-calibration and the room inference tasks which are again separated in general, with the self-calibration block yielding the estimated positions $\hat{\mathbf{r}}$ and $\hat{\mathbf{s}}$ of the microphones and of the calibration sources, using only the direct sound parameters. These positions are then needed to correctly obtain the reflector parameters and to position the reflectors in the final scene model, i.e. for performing the room inference.

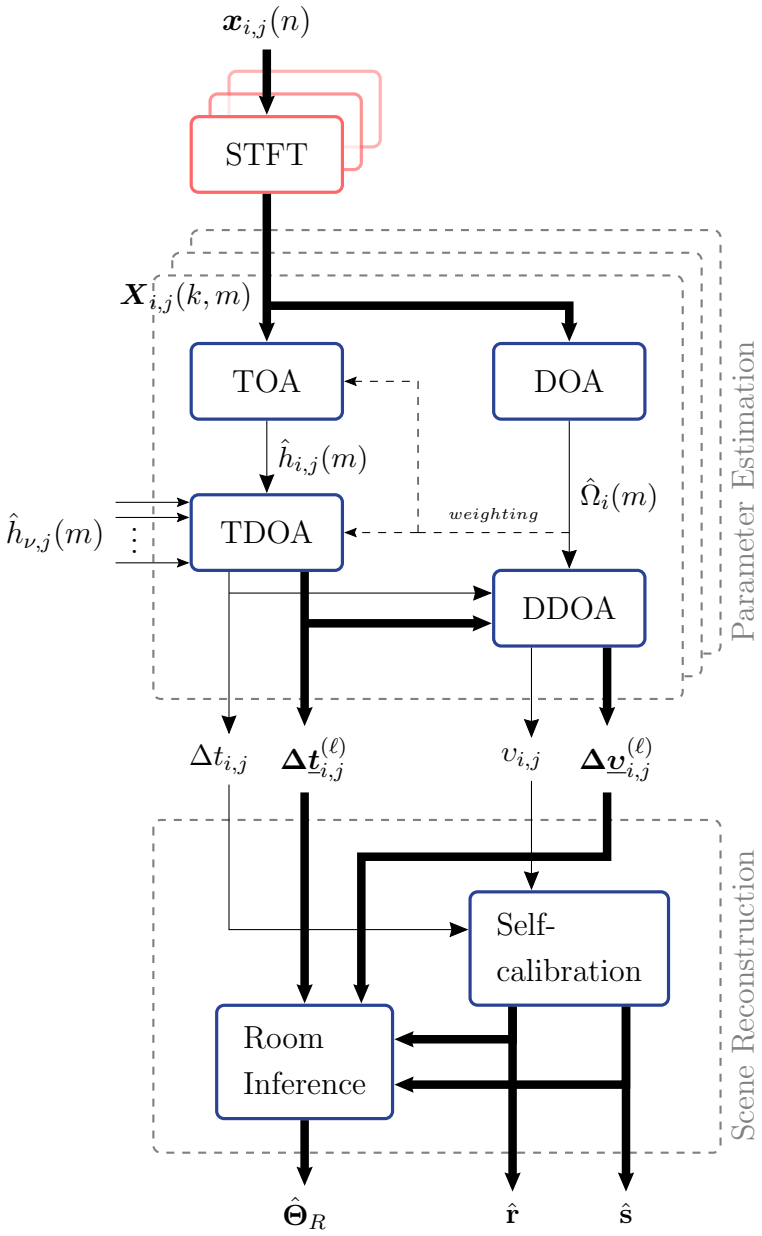


Figure 6: Flow graph showing the essential signal processing blocks of the parameter estimation and scene reconstruction algorithms, indicated by **dashed grey** rectangles. The parameter estimation as well as the STFT of the microphone signals are performed for all microphones, which are interacting with each other by means of the TOA functions $\hat{h}_{\nu,j}(m)$ and $\hat{h}_{i,j}(m)$ representing a pseudo-impulse response for each microphone i and $\nu = \{1, 2, \ldots, N_i\} \setminus i$. $\hat{\Omega}_i(m)$ represents a DOA function giving an estimated DOA value at each frame m.

1.2.2 Thesis Outline

The thesis will be organized as follows:

Chapter 2 describes the implemented algorithms for the TOA and DOA estimation tasks.

Chapter 3 describes the framework for performing self-calibration and room inference from the parameters estimated in Chapter 2.

Chapter 4 shows results for scene reconstruction performed on data obtained by measurements in two different rooms.

Chapter 5 gives a summary of all results as well as application examples and possibilities of future expansions of the algorithms.

2 Parameter Estimation

This section gives a short overview over *time of arrival* (TOA) and *direction of arrival* (DOA) estimation algorithms in Section 2.1 followed by the proposed algorithms to estimate the parameters needed for the self-calibration and room inference tasks. These parameters are the DOAs of direct sound and reflections and the TOAs of direct sound and reflections, with the proposed DOA estimator described in Section 2.2 and the proposed TOA estimator in Section 2.3.

2.1 Literature Review

Considering that a large number of DOA estimation algorithms rely in some form on timing information, TOA estimation and the resulting knowledge of TDOAs can be seen as the backbone of DOA estimation and acoustic scene reconstruction.

A thorough overview of the popular algorithms for estimating TDOAs between signal is given by Chen et al. in [CBH06] (there termed *time delay estimation* (TDE) describing the act of TDOA estimation), covering basic algorithms which use only two microphones and extensions to the corresponding multichannel setups. An important algorithm group described therein is made up of a number of cross correlation based ones, for example simple *cross correlation*, *generalized cross correlation* and various extensions thereof, all of which maximize a function derived from the cross correlation function. Further examples are the *adaptive eigenvalue decomposition* and adaptive TDE algorithms (usually using some form of least mean squares algorithm in dual or multichannel form), both trying to estimate the *room impulse response* (RIR) from the source to the respective microphones. The time differences between the maxima in the respective RIRs (assumed to correspond to the direct path) are then used as an estimate for the time delay between the signals at the examined microphones. An extension for the dual channel method is given in the form of a fusion algorithm using multiple sensor pairs and allowing a combination of different cost functions.

Correlation based algorithms can be found in the works by Knapp and Carter [KC76] describing the generalized cross correlation method, by Omologo et al. [OS94] using the crosspower-spectrum phase for TDE and by Nesta et al. in [NOS08a] and [NOS08b] describing an extension of generalized cross correlation techniques suitable for separating and finding TDOAs of two or more sources as well as for

localizing said sources in highly reverberant environments using short, continuous source signals.

Rather similar to the correlation based ones, TOA algorithms exist in the field of template matching as well. An example is *dynamic time warping* which was introduced by Sakoe and Chiba [SC78] for use in speech recognition and adapted by Kelly and Boland [KB14] for finding reflections in RIRs. Dynamic time warping allows stretching, shifting and scaling of an acoustic template, resulting in an overall nonlinear transformation which makes it interesting for TOA estimation. Another template matching based algorithm is described by Dokmanic and Vetterli in [DV15] as a modification of the *orthogonal matching pursuit*. Here a dictionary of orthogonal functions is derived from the received sensor signals (from which the TOAs can be retrieved) by use of a template of an emitted probing pulse, also yielding the filters to create the respective dictionary entries (i.e. the filters corresponding to reflective surfaces). Template matching algorithms can obviously be used on a single microphone as well as on multiple microphones.

Apart from the possibility of using a single microphone array for time based DOA estimation, the use of distributed microphone arrays has the advantage of allowing space-time signal processing, fusing together the data collected by each array at a different spatial location. A good overview of array signal processing is given by Krim and Viberg in [KV96], alongside an overview of DOA estimators, representing one of the most important tasks of microphone arrays in general. DOA estimates are often used for source localization (especially when using unsynchronized microphone arrays) for which knowledge of the microphone positions is needed as prior information.

Apart from time based algorithms also subspace processing algorithms for DOA estimation are very popular, working on arbitrary (but known) array geometries and performing a separation of the recorded signal into the signal and noise subspace. Possible DOAs are found at the intersection of the signal subspace and the array manifold (i.e. the set of all array steering vectors). The most popular algorithms in this class are MUSIC[2] by Schmidt [Sch86] and ESPRIT[3] by Roy [RK89], which are easily translated into the spherical harmonics domain, shown for example by Li [LYMH11] or by Sun et al. [STMK11]. There also exist a large number of extensions [DW86, ZKS93, Kun96, ML99].

Another popular class are beamformer based approaches, which aim at maximizing the output power of an array to find possible DOAs. Two members of this class are

[2]Multiple Signal Classification

[3]Estimation of Signal Parameters via Rotational Invariance Techniques

the Bartlett beamformer, which uses a constraint on filter weights, or the Capon beamformer, which calls for no distortion in a certain direction and is popularly known as *minimum variance distortionless response* beamformer. Both of these are described in [KV96]. Another example belonging to this group is the *informed linearly constrained minimum variance* beamformer described by Thiergart and Habets [TH13], where multiple DOAs are estimated by means of spatial filtering, minimizing the diffuse and self-noise power.

Examples for instantaneous DOA estimation methods to compute a vector that represents an instantaneous estimate of the acoustic intensity at each time frame are described by Williams et al. [WVHK06] and Pavlidi [PDMPM15]. Another instantaneous DOA estimator is proposed by Politis et al. [PDMP15] by a direct weighting of each sensor look direction with the recorded signal (performed in the frequency domain), shown to be applicable to spherical arrays that fulfill certain properties.

2.2 Direction of Arrival

From the DOA estimators mentioned beforehand, the estimators performing instantaneous estimates have the benefit of offering a fast and simple way to acquire DOA estimates at each time instance. Despite this advantage, these types of estimators are usually restricted to a certain array dependent frequency range. When using spherical microphone arrays, the upper frequency bound (aliasing frequency) is set by the number of microphones used to sample the sphere as well as the radius at which the capsules are located. The lower bound is given by not ideal capsule characteristics towards low frequencies. For the *Oktava* microphones used in this work the usable frequency range would be limited to roughly 500 to 2500 Hz which would leave only very few frequency bins to work with when attempting to acquire instantaneous DOA estimates for each time-frequency bin.

Politis et al. [PDMP15] propose an instantaneous DOA estimator that overcomes this frequency restriction. A broadband instantaneous DOA estimator would be very useful when performing the DOA estimation in the frequency domain, exploiting the broadband character of the calibration signals as well as the fact that at every time-frequency bin only a single source is assumed to be active (which is true for the direct sound and mostly true for reflections), allowing the use of all frequency bins at each time instance without the need to assign frequencies to different sources. The available number of frequency bins is then only restricted by the FFT size, bearing in mind that as the arrivals of direct sound and reflections might be closely spaced,

the main focus should be on a good time resolution (i.e. a short window length and a reasonable FFT length).

The algorithm proposed by Politis et al. [PDMP15] which was taken as a starting point for improvements is described briefly in Section 2.2.1 and the proposed extension described in Section 2.2.2 thereafter.

It should be noted that beamformer based DOA estimators would not suffer from the frequency bound imposed by aliasing but in turn require searching over the whole interesting direction range.

2.2.1 DOA Estimation based on Magnitude Sensor Response

Politis et al. [PDMP15] propose a DOA estimator by evaluating the superposition of the look directions of the capsules of a spherical microphone array weighted with the magnitude of the recorded signals. The resulting vector then points into the direction of the estimated DOA, allowing estimates above the aliasing frequency. The underlying assumption is that the used microphone array allows the integration over all magnitude distributions of the capsules to a non-zero value (the capsules should be directional and exhibit an identical symmetric pattern), according to the formula given by Politis et al. as

$$\hat{\boldsymbol{y}}_{\mathrm{DOA}}(k,t,\Omega_{\mathrm{dir}}) = |S_{\mathrm{dir}}(k,t)| \int_{\Omega} |h(k,R,\alpha)|\mathbf{n}(\Omega)d\Omega. \tag{2.2.1}$$

In upper equation $S_{\mathrm{dir}}(k,t)$ is the direct sound signal, $|h(k,R,\alpha)|$ is the magnitude distribution of the microphone array, α the angle between the impinging wave and the measurement point and $\mathbf{n}(\Omega)$ the unit vector in direction Ω. R is the radius of the sphere sampled by the array capsules. The vector $\hat{\boldsymbol{y}}_{\mathrm{DOA}}(k,t,\Omega_{\mathrm{dir}})$ then points in the direction of the impinging plane wave. The constraints on the array geometry are described in detail in [PDMP15]. The discretized version of Equation 2.2.1 described by Politis et al. will now be described, applied to the microphone array used in this work.

The look directions of the microphone capsules in each microphone array can be described by the unit direction vector in Cartesian coordinates

$$\boldsymbol{\nu}_{\kappa} = \begin{pmatrix} x_{\kappa} & y_{\kappa} & z_{\kappa} \end{pmatrix}^{T}, \tag{2.2.2}$$

relative to the centre of each microphone $\mathbf{r}_i$ (see Figure 7), where $\kappa \in \{1,2,3,4\}$ is the index for the different capsules. These capsule vectors $\boldsymbol{\nu}_{\kappa}$ are assumed to be

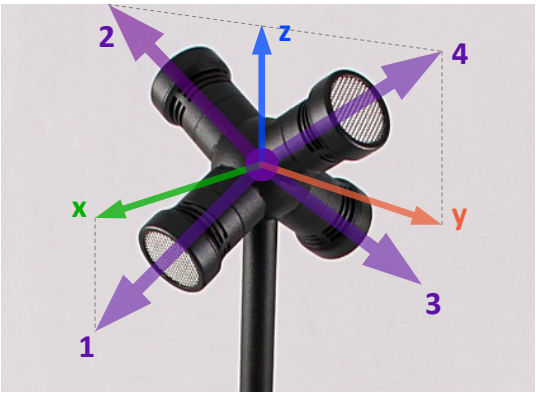


Figure 7: Capsule vectors representing the look direction of each microphone capsule in the local coordinate system. The picture is taken from [Okt], the coordinate system is included for clarification.

the same when using microphone arrays of the same type. The first capsules look direction lies in the xz-plane, with the numeration continuing in a clockwise fashion when projected onto the xy-plane. The individual vectors are combined in a matrix as

$$\mathbf{N}_K = \begin{pmatrix} | & | & | & | \\ \boldsymbol{\nu}_1 & \boldsymbol{\nu}_2 & \boldsymbol{\nu}_3 & \boldsymbol{\nu}_4 \\ | & | & | & | \end{pmatrix}. \tag{2.2.3}$$

The DOA estimation performed in the frequency domain then results in a direction estimate at each time-frequency bin of the recorded signal.

The direct weighting of the look directions with the respective signals in the frequency domain can be computed as

$$\hat{\boldsymbol{y}}_{i,j}(k,m) = \mathbf{N}_K \cdot |\boldsymbol{X}_{i,j}(k,m)|, \tag{2.2.4}$$

where the vector $\hat{\boldsymbol{y}}_{i,j}$ is the estimated DOA in Cartesian coordinates at each frequency bin k and time frame m for microphone i. $\boldsymbol{X}_{i,j}(k,m)$ is a column vector of size (4×1) containing the STFT of the capsule signals from microphone i at frequency k and time frame m. The index j indicates which source is active in the currently evaluated signal block. Each capsules direction is weighted by the magnitude of the recorded signal spectrum, resulting in a direction estimate at every frequency and time frame.

It should be noted that Politis et al. use this DOA estimator for frequencies above the aliasing frequency of the array and *intensity-based* DOA estimation below the aliasing frequency.

2.2.2 Smoothed Magnitude Response

This section describes the proposed enhancement of the magnitude response algorithm from [PDMP15], which is furthermore used on the whole frequency range due to the satisfying results.

The proposed extension of the magnitude sensor response algorithm is to perform a separation into signal and noise subspace (similar to the MUSIC approach), aiming at improving the direction estimates in case of noise by using the eigenvector of the largest eigenvalue of the covariance matrix at each frequency as a weight for the capsule vectors (instead of the magnitude of the recorded signal spectrum). The computations are performed for all microphones separately, thus where obvious the indices i and j denoting the microphone array and active calibration source are dropped for readability.

An estimate $\hat{\mathbf{\Sigma}}_{i,j}(k,m)$ of the covariance matrix $\mathbf{\Sigma}_{i,j}(k,m)$ of size $(K \times K)$, with i denoting the microphone index, j the source index and K the number of microphone capsules, is computed on a window of M frames around the current time frame m according to

$$\hat{\mathbf{\Sigma}}_{i,j}(k,m) = \frac{1}{M} \sum_{\mu=m-M/2}^{m+M/2-1} \boldsymbol{X}_{i,j}(k,\mu) \cdot \boldsymbol{X}_{i,j}(k,\mu)^H, \tag{2.2.5}$$

with the number of frames over which it is computed kept short (e.g. $M = 32$) to maintain a good separation between closely spaced reflections and to ensure that the assumption that only a single signal is active us fulfilled most of the time. This estimated Covariance matrix can be decomposed into

$$\hat{\mathbf{\Sigma}} = \mathbf{U}\mathbf{\Lambda}\mathbf{W} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^H, \tag{2.2.6}$$

due to symmetry. $\mathbf{U}$ represents a (4×4)-matrix containing the left eigenvectors, $\mathbf{\Lambda}$ a diagonal matrix containing the eigenvalues λ_κ in decreasing order on the main diagonal, and $\mathbf{W}$ contains the right eigenvectors. This decomposition is performed at all time-frequency bins.

Under the assumption that only one signal is present at any time frame m, the largest eigenvalue and corresponding eigenvector span the signal subspace $\lambda_S = \lambda_1 = \sigma_S^2$, with the rest of the eigenvalues and eigenvectors corresponding to the noise subspace $\lambda_N \geq \lambda_2 \geq \lambda_3 \geq \lambda_4 = \sigma_N^2$ which are all equal to the noise variance in the ideal case[4], but in reality usually somewhere close by. The separation into signal and noise

[4]The ideal case being that there *actually is* only a single signal source present and that the

subspace yields

$$\hat{\mathbf{\Sigma}} = \mathbf{U}\mathbf{\Lambda}_S\mathbf{U}^H + \mathbf{U}\mathbf{\Lambda}_N\mathbf{U}^H \tag{2.2.7}$$

with $\mathbf{\Lambda}_S = \text{diag}\{\lambda_1, 0, 0, 0\}$ and $\mathbf{\Lambda}_N = \text{diag}\{0, \lambda_2, \lambda_3, \lambda_4\}$

The signal subspace can be mapped back to microphone capsule directions at each time-frequency bin by evaluating

$$\tilde{\boldsymbol{y}}(k, m) = \mathbf{N}_K \cdot |\mathbf{u}_S(k, m)|, \tag{2.2.8}$$

with $\mathbf{u}_S(k)$ representing the eigenvector to λ_S. The resulting DOA estimates $\tilde{\boldsymbol{y}}_{i,j}(k, m)$ in Cartesian coordinates (again including the indices for microphone i and source j) at frequency bin k and time frame m can then be written in spherical coordinates as

$$\tilde{\boldsymbol{y}}^{\circ}_{i,j}(k, m) = \begin{pmatrix} r_{i,j}(k, m) & \phi_{i,j}(k, m) & \theta_{i,j}(k, m) \end{pmatrix}^T. \tag{2.2.9}$$

The following computations that describe the computation of the final DOA estimates can be applied on the azimuth and elevation angle in the same way, resulting in the respective azimuth and elevation DOA estimates. The procedure will be described using the azimuth angle of $\tilde{\boldsymbol{y}}^{\circ}_{i,j}$, i.e. using

$$\upsilon_{i,j}(k, m) = \phi_{i,j}(k, m). \tag{2.2.10}$$

The DOAs $\upsilon_{i,j}(k, m)$ are then examined by performing a histogram over all frequencies k at each frame m to find the number of votes a direction got, computed according to

$$q_{i,j}(m, b) = \sum_k \Pi_{b,\Delta_b}\left(\upsilon_{i,j}(k, m)\right) \tag{2.2.11}$$

with

$$\Pi_{n,N}(x) = \begin{cases} 1, & N \cdot \left(n - \frac{1}{2}\right) < x \leq N \cdot \left(n + \frac{1}{2}\right) \\ 0, & else \end{cases} \tag{2.2.12}$$

where $b \in \{1, 2, \ldots, N_b\}$ is the index of the histogram bin, N_b is the number of histogram bins and Δ_b the width of the histogram edges computed with $\Delta_b = \frac{2\pi}{N_b}$

estimate of the covariance matrix is perfect. The interested reader is referred to the works of Mestre et al. [Mes08, ML08] and Yazdian et al. [YGB13] for thorough investigations of the asymptotic behaviour of the sample covariance matrix.

for the azimuth angle. Π_{b,Δ_b} counts the number of all frequencies for which the estimated DOA lies within that very histogram interval. The histogram data is then stored in vector and matrix form according to

$$\mathbf{q}_{i,j}(m) = \Big(q_{i,j}(m,1) \quad q_{i,j}(m,2) \quad \cdots \quad q_{i,j}(m,b) \quad \cdots \quad q_{i,j}(m,N_b) \Big)^T \tag{2.2.13}$$

$$\mathbf{Q}_{i,j} = \Big(\mathbf{q}_{i,j}(1) \quad \mathbf{q}_{i,j}(2) \quad \cdots \quad \mathbf{q}_{i,j}(m) \quad \cdots \quad \mathbf{q}_{i,j}(N_m) \Big) \tag{2.2.14}$$

where N_m is the number of frames that are analyzed.

The DOA estimate at a certain time frame m is then the location of the maximum in the respective column, computed using

$$\hat{b}(m) = \underset{b}{\operatorname{argmax}} \left\{ q_{i,j}(m,b) \right\}. \tag{2.2.15}$$

Since the DOA histograms might not exhibit a single peak symmetrical around the real value (caused by the use of finite histogram bins) a slight shift towards the centre of mass around the found maximum $\hat{b}(m)$ is allowed,

$$\hat{v}_{i,j}(m) = \frac{\sum_{\beta=\hat{b}(m)-\eta}^{\hat{b}(m)+\eta} \beta \cdot q_{i,j}(m,\beta)}{\sum_{\beta=\hat{b}(m)-\eta}^{\hat{b}(m)+\eta} q_{i,j}(m,\beta)}, \tag{2.2.16}$$

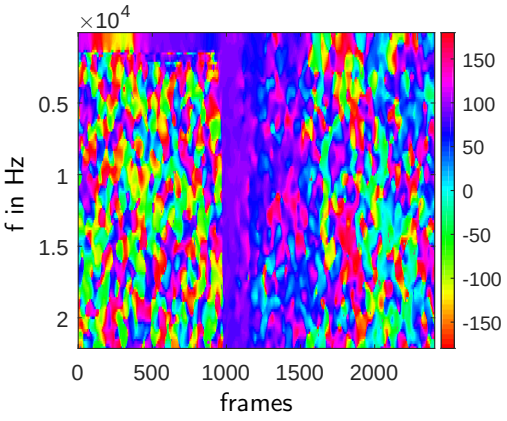
where η is the number of bins around the global maximum that are included in the evaluation (chosen with $\eta < 10$ with $N_h = 90$ overall histogram bins), removing the discretization of the possible DOAs introduced by the histograms.

2.2.3 DOA Estimation Results

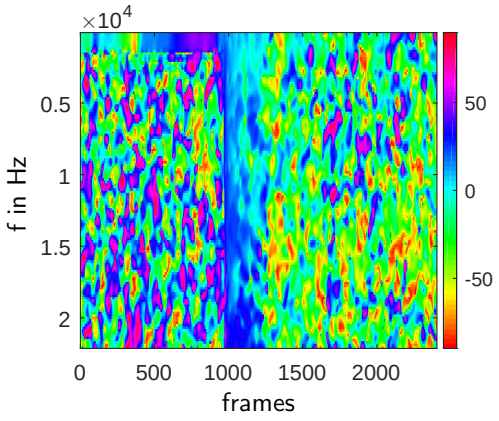
The DOA estimation results in azimuth and elevation angles for two data sets taken from the first measurement (described in Appendix B.1) can be seen in Figure 8. The DOA spectrum $\tilde{\boldsymbol{y}}_{i,j}(k,m)$ shows the estimated DOA in degrees at each frequency and time frame over an examination window of 2500 *samples*. For computing the STFT a short Hann window of length $N_{\text{win}} = 64$ *samples*, an FFT length of $N_{\text{FFT}} = 256$ and a hop size $N_{\text{hop}} = 1$ *samples* are used. For computing the estimate of the covariance matrix at each frequency a normalized rectangular window of length $N_{\text{cov}} = 32$ *samples* is used. These parameters showed the most promising results during initial development, allowing good time resolution due to the short time window, and a moderate number of frequency bins, though at a rather low frequency resolution. This low resolution poses no problem since the recorded claps should still be sufficiently broadband and examining specific frequencies is not of interest.

For the azimuth DOA in Figures 8a and 8c the arrival of the direct sounds between frame indices $1000 - 1500$ is clearly visible as a band of similar DOA estimates over the whole frequency range. For the elevation DOA estimate the direct sound (which should be at $\theta = 0°$) is visible less clearly, observable in Figures 8b and 8d.

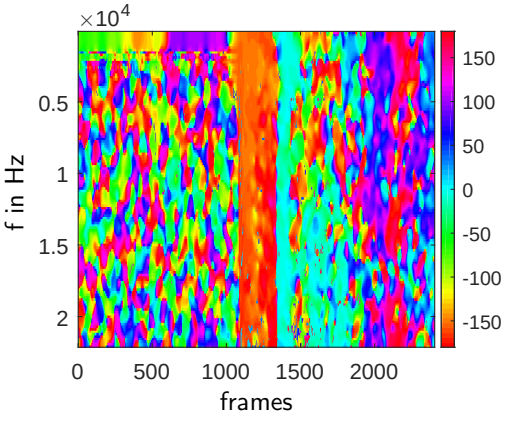
Examples for the frame wise histograms $\mathbf{Q}_{i,j}$ are displayed in Figure 9, 10 and 11 in the respective middle plots as colour plots showing the frame index along the x- and the estimated DOA in degrees along the y-axis.



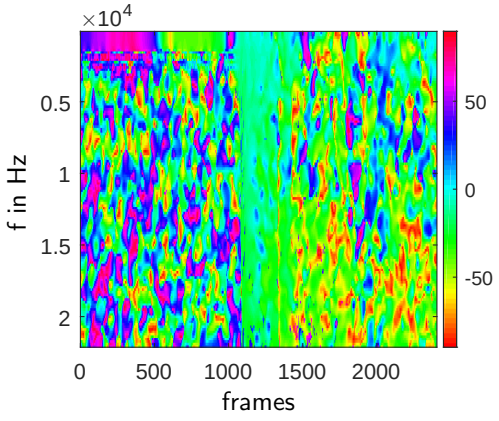
(a) $i = 1$, $j = 1$, azimuth DOA spectrum



(b) $i = 1$, $j = 1$, elevation DOA spectrum



(c) $i = 1$, $j = 2$, azimuth DOA spectrum



(d) $i = 1$, $j = 2$, elevation DOA spectrum

Figure 8: DOA spectra $\tilde{\boldsymbol{y}}_{i,j}(k, m)$ for azimuth (left) and elevation (right) angles for two sources recorded by the same microphone.

2.3 Time of Arrival

The next parameters to assess are the TOAs of direct sounds and reflections that are needed to determine the TDOAs of direct sound and reflections within the signals recorded by the microphones. The first arrival is assumed to match the direct sound and all later arrivals to reflections.

2.3.1 Eigenvalue based TOA Estimation

Apart from the TOA estimators mentioned in Section 2.1, an interesting approach is described by Zeng [YC08], describing how the eigenvalues of the covariance matrix of a received signal can be used to detect the presence of signals in cognitive radio. Assuming that a similar assumption is applicable in acoustics, the arrival of direct sound or a reflection can be compared to the activity of a user in a radio channel. Since the eigenvalues are already computed for the DOA estimation algorithm, it seems obvious that the amplitude of the largest eigenvalue shows some form of connection to the arrival of a signal, either direct sound or reflection.

This section proposes a simple method to find TOA candidates for the direct sound and reflections, again using histograms, since the eigenvalues of the covariance matrix are already available for the whole frequency range, using (optional) weights that are assumed to correspond to the reliability of the DOA estimates at each time frame m. As practically all existing DOA algorithms output some form of reliability measure for possible DOA angles, it should be possible to find a suitable weight for other DOA algorithms as well.

Frequency Domain Eigenvalue Picking
With the amplitude of the largest eigenvalue of the sample covariance matrix at each frequency k and time frame m available as $\lambda_S(k, m)$, initial TOAs for each frequency can be found by picking N_p peaks over the temporal evolution of the eigenvalues at each frequency. The notion behind this is that even though the claps are usually much longer than a single sample (usually something around 200 *samples* as can be assumed Figure 55) they can still be assumed broadband. These percussive 'ridges' should ideally span over a large portion of frequencies at the TOAs of direct sound and reflections containing large enough energy. TOAs are then found by evaluating all peaks found at all frequencies using histograms.

A possible way to improve the percussive character of the signals can be found in the work by Fitzgerald [Fit10], who describes the separation of audio signals into

percussive and harmonic parts by creating spectral masks from the median filtered signal spectrum. The filtering is performed over frequency to create a percussive mask and over time for the harmonic mask. Fitzgerald then compares these masks at each time-frequency bin for assigning the respective bins of the signal spectrum to the harmonic or percussive spectrum. Since only the percussive part is of interest here, only the percussion enhancing median filtering will be performed (no comparison of masks is performed).

The percussion enhanced signal eigenvalue spectrum at time frame m is computed according to

$$\hat{\lambda}_S(k,m) = \text{median}_k \left\{ \lambda_S(k,m) \right\}, \tag{2.3.1}$$

using a moving median filter over frequency bins k with the filter length L_{median} as the only parameter. The used implementation of the moving median shortens the filter when moving closer to the edges of $\lambda_S(k,m)$, resulting in no size reduction of $\hat{\lambda}_S(k,m)$.

On this percussion enhanced eigenvalue magnitude spectrum, peak picking is performed following

$$\boldsymbol{\ell}_\lambda(k) = \underset{N_p \;\rightarrow\; m}{\text{PP}} \left\{ \hat{\lambda}_S(k,m) \right\}, \tag{2.3.2}$$

where the operator $\underset{N_p \;\rightarrow\; m}{\text{PP}} \{x(m)\}$ picks the N_p locations of the largest local maxima of the signal $x(m)$ over all available time frames m, storing the locations in a $(N_p \times 1)$-vector. The peak locations in the vector are sorted according to the magnitude of the peak. The output vector $\boldsymbol{\ell}_\lambda(k) = \begin{pmatrix} \ell_\lambda(k,1) & \cdots & \ell_\lambda(k,p) & \cdots & \ell_\lambda(k,N_p) \end{pmatrix}^T$ contains the TOA candidates for each frequency k and labeled with p.

TOA and DOA Fusion

After the peak picking stage, a weighted histogram is created using the peaks $\boldsymbol{\ell}_\lambda(k)$ of all frequencies k (assuming that the excitation signal is reasonably broadband). For the histogram weights $w_t(m)$ the logarithmic magnitude of the signal space eigenvalues $\hat{\lambda}_S(k,m)$ and a measure for the density of the DOA estimation at the respective time-frequency bin are combined.

The density measure $\rho(x)$ is taken from Du [DDJ16] (used therein to find possible initializations for a clustering algorithm) and corresponds to locations where points have a low distance to a large number of other points.

Since the DOAs are only available as polar angles a distance will be computed using complex phasors, defining the distance between two DOAs in radians (with the usual indices i, j, and m omitted for readability) as

$$d\big(\upsilon(k), \upsilon(\kappa)\big) = \left|e^{i\cdot\upsilon(k)} - e^{i\cdot\upsilon(\kappa)}\right|, \tag{2.3.3}$$

where k and κ represent two frequency bins and $\upsilon(k)$ a DOA estimate for frequency k. This allows the DOA angles to wrap around 2π. With the distance between two angles defined, the computation of the density is given by Du [DDJ16] in two ways: either by using a hard threshold and counting the points within the threshold distance d_c, as

$$\rho(k) = \sum_{\kappa} \chi\left(d\big(\upsilon(k), \upsilon(\kappa)\big) - d_c\right) \tag{2.3.4}$$

$$\chi(x) = \begin{cases} 1, & x < 0 \\ 0, & x \geq 0 \end{cases}$$

or by using a soft threshold of form

$$\rho(k) = \sum_{\kappa} \exp\left(-\frac{d\big(\upsilon(k), \upsilon(\kappa)\big)^2}{d_c^2}\right) \tag{2.3.5}$$

with the hard threshold used in this work.

Because for the case examined here the frequency where the maximum of the density occurs is not of interest, only the maximum value of the density at each time frame is used as

$$\rho_{\max}(m) = \max_{k} \rho(k, m), \tag{2.3.6}$$

indicating the similarity between each frequencies DOA estimates at each time frame m.

To find the weights for the TOA histograms of an examined signal block that contains most of the RIR (chosen depending on the room size), the percussion enhanced amplitude spectrum of the signal eigenvalue $\hat{\lambda}_S(k, m)$ is normalized such that $\bar{\lambda}_S(m, k) \in [0, 2]$ and the density $\rho_{\max}(m)$ is normalized to a range of $\bar{\rho}_{\max}(m) \in [-1, 1]$ in the examined signal block. The final weight is then computed as

$$w_t(k, m) = u\big(\bar{\lambda}_S(k, m) + \bar{\rho}_{\max}(m)\big) \tag{2.3.7}$$

where $u(x)$ is a function defined as

$$u(x) = \begin{cases} x, & x > 0 \\ 0, & x \leq 0 \end{cases} \tag{2.3.8}$$

which sets all negative values to zero.

This allows for the weights $w_t(k, m)$ to be reduced by the usually low density at frames with no real peaks in the DOA histograms (i.e. negative normalized density $\bar{\rho}_{\max}$), as well as increased when the DOA histogram at frame m shows a sharper peak (i.e. positive normalized density $\bar{\rho}_{\max}$).

The weighted histogram of the peaks picked at all frequencies k, which can be called a pseudo-*room impulse response* (pseudo-RIR), is then computed as

$$h_(m) = \sum_{k} \sum_{p=1}^{N_p} w_t(k, \ell_\lambda(k, p)) \cdot \Pi_{m,\Delta_m} \left(\ell_\lambda(k, p)\right), \tag{2.3.9}$$

with $\ell_\lambda(k, p)$ representing the p-th peak at frequency k and Δ_m the histogram bin width, using again

$$\Pi_{n,N}(x) = \begin{cases} 1, & N \cdot \left(n - \frac{1}{2}\right) < x \leq N \cdot \left(n + \frac{1}{2}\right) \\ 0, & else. \end{cases} \tag{2.3.10}$$

Equation 2.3.9 is evaluated using the parameters obtained by the respective microphone and source combination, resulting in the pseudo-RIRs $h_{i,j}(m)$ for microphone i and source j .

Plots of resulting pseudo-RIRs $h_{i,j}(m)$ can be seen in Figures 9, 10 and 11 in the bottom plots of each subplot. The middle plots show the DOA histograms and the respective top plots the DOA spectra (the estimated DOAs at each frequency). The results are plotted for the three microphones and the clap types *broadband* (bb), *low pass* (lp) and *high pass* (hp) for the source position B (or $j = 2$) with measurements obtained from the measurement room (described in Appendix B.1). The type is assigned to the claps by auditory inspection and to indicate that different claps are created intentionally.

The pseudo-RIRs $h_{i,j}(m)$ are then again the basis for picking N_t peaks to find the TOAs of direct sound and reflections, with the earliest peak found expected to match the direct sound TOA and all others reflections. The TOA candidates for microphone i of the active source j are picked from the pseudo-RIRs and stored in a vector

according to

$$\bar{\mathbf{t}}_{i,j} = \underset{N_t \;\rightarrow\; m}{\mathrm{PP}} \left\{ h_{i,j}(m) \right\}, \tag{2.3.11}$$

for microphone i and source j. The TOA vector has the form

$$\bar{\mathbf{t}}_{i,j} = \begin{pmatrix} t_{i,j}(1) & \cdots & t_{i,j}(\tau) & \cdots & t_{i,j}(N_t) \end{pmatrix}^T \tag{2.3.12}$$

with $\tau \in \{1, 2, \ldots, N_t\}$ labelling the peaks. The direct sound TOA $t_{i,j}$ is then assumed as the earliest TOA at each microphone array computed by

$$t_{i,j} = \min \; \bar{\mathbf{t}}_{i,j} \tag{2.3.13}$$

with minimum operator applied to a vector returning the smallest entry of that vector. From these TOAs the direct sound TDOAs for source j can be computed with the TOAs at all microphones i as

$$\Delta t_{i,j}^{(i_0)} = t_{i,j} - t_{i_0,j}, \tag{2.3.14}$$

where $t_{i_0,j}$ is the reference TOA for source j, i.e. the earliest arrival at any microphone array. This reference array indexed with i_0 is computed for each source j as

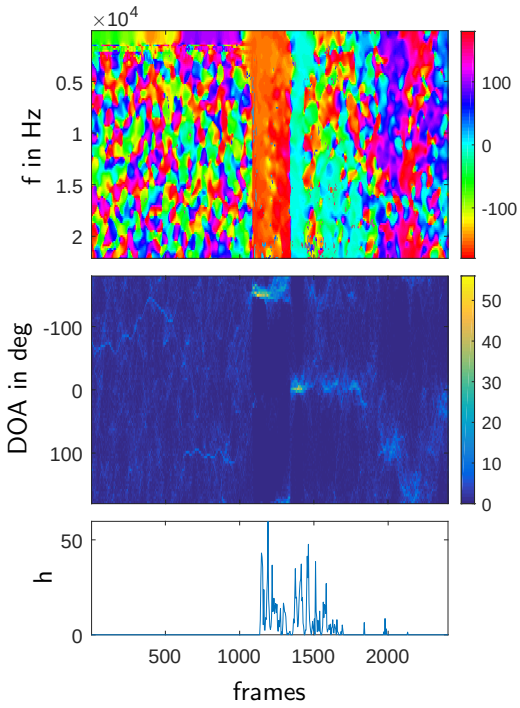
$$i_0(j) = \underset{i}{\mathrm{argmin}} \left\{ t_{i,j} \right\} \tag{2.3.15}$$

The reflection TDOAs are computed from the picked TOAs $t_{i,j}(\tau)$ of a single array as

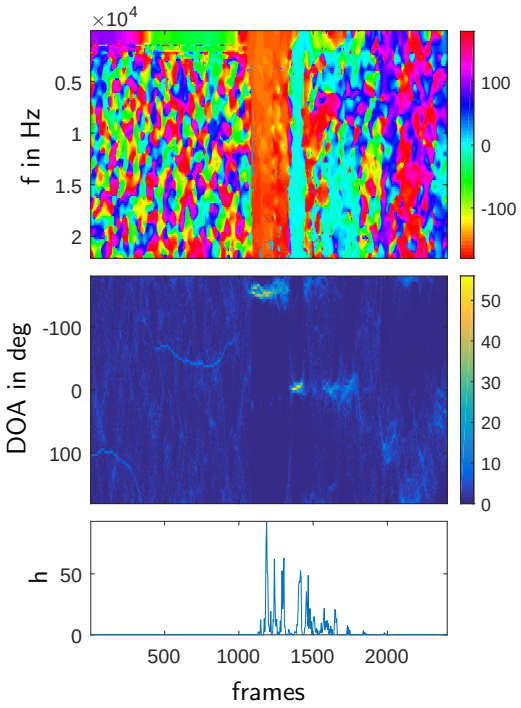
$$\Delta \underline{t}_{i,j}^{(\ell)} = t_{i,j}(\ell) - t_{i,j} \tag{2.3.16}$$

for each peak picked from $h_{i,j}(m)$ where ℓ is used to index all remaining $N_t - 1$ peak positions which are those assumed to correspond to reflections.

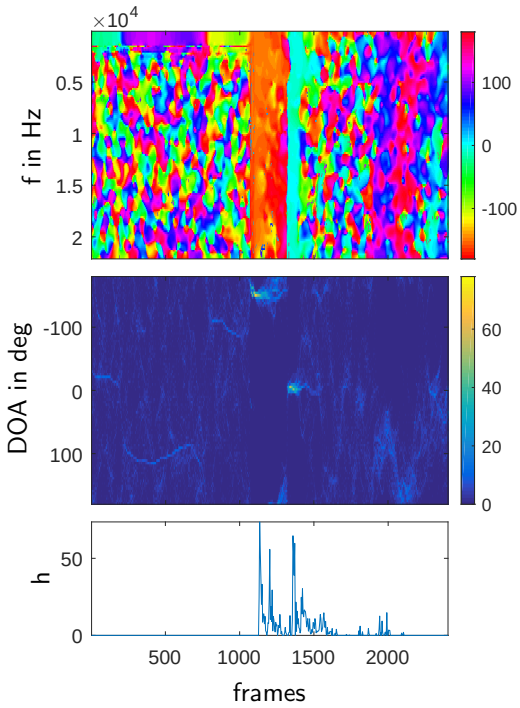
It should be noted that the use of the STFT with a specific hop size N_{hop} and histograms with a certain bin spacing Δ_b affect the resulting resolution of the TOA estimation.



(a) $i = 1$, $j = 2$, bb

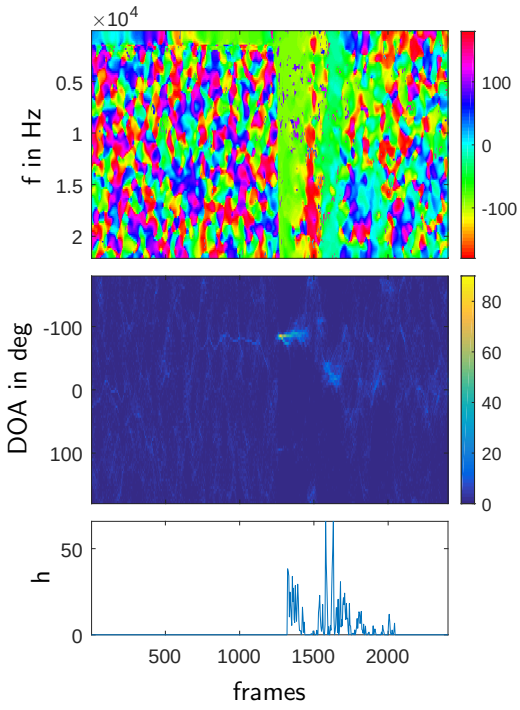


(b) $i = 1$, $j = 2$, lp

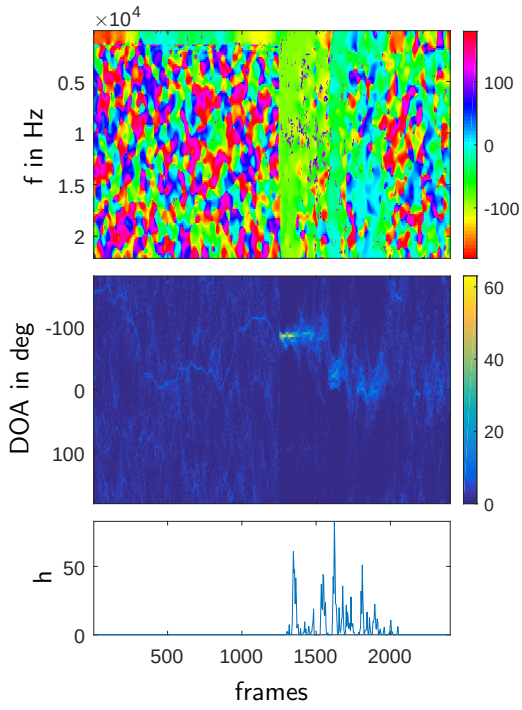


(c) $i = 1$, $j = 2$, hp

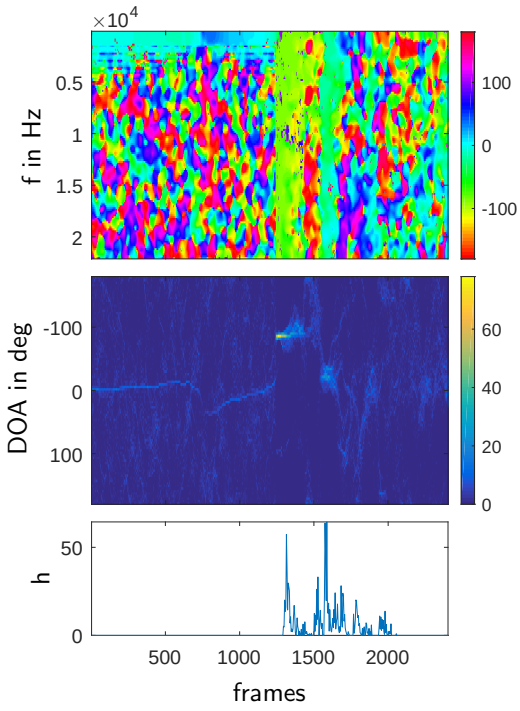
Figure 9: The subplots show the DOA spectra in degrees (*top*), DOA histograms (*middle*) and pseudo-RIRs $h_{i,j}(m)$ (*bottom*). Results are shown for three clap types for a microphone source combination.



(a) $i = 2$, $j = 2$, bb

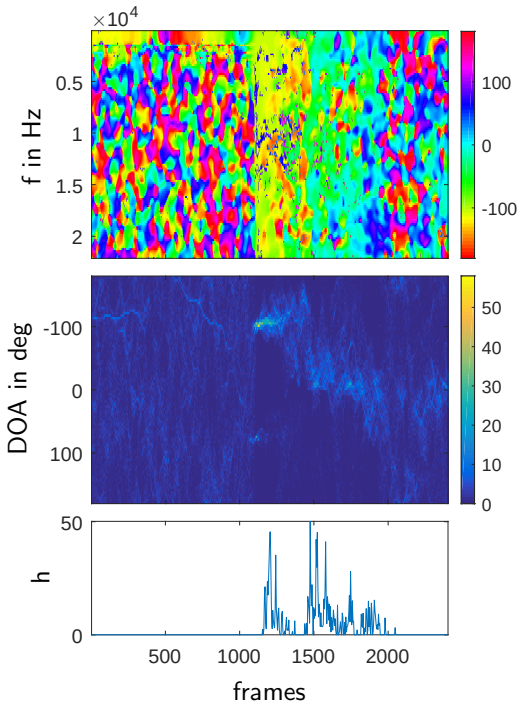


(b) $i = 2$, $j = 2$, lp

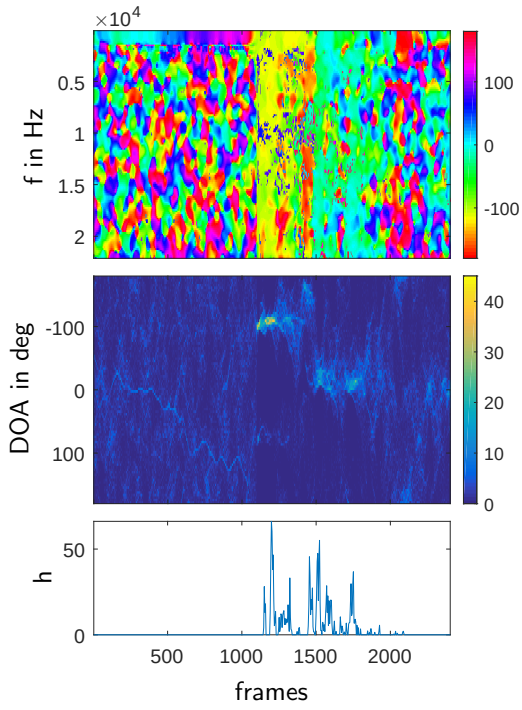


(c) $i = 2$, $j = 2$, hp

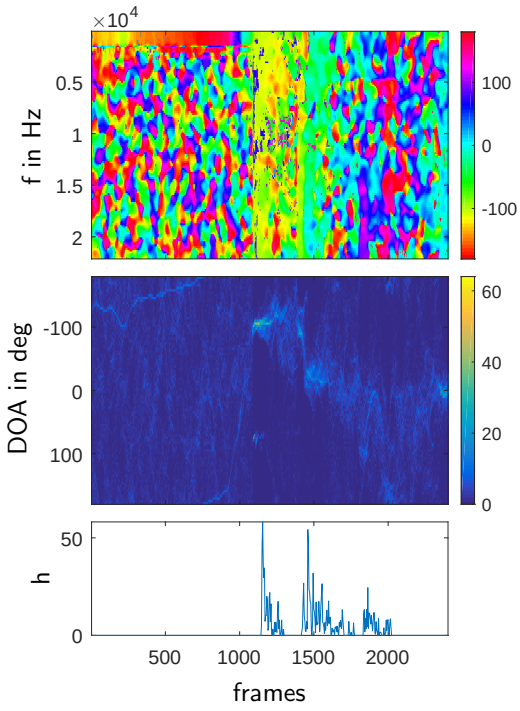
Figure 10: The subplots show the DOA spectra in degrees (*top*), DOA histograms (*middle*) and pseudo-RIRs $h_{i,j}(m)$ (*bottom*). Results are shown for three clap types for a microphone source combination.



(a) $i = 3$, $j = 2$, bb



(b) $i = 3$, $j = 2$, lp



(c) $i = 3$, $j = 2$, hp

Figure 11: The subplots show the DOA spectra in degrees (*top*), DOA histograms (*middle*) and pseudo-RIRs $h_{i,j}(m)$ (*bottom*). Results are shown for three clap types for a microphone source combination.

2.4 Parameter Estimation Results

The results for the TOA and DOA estimation performed on data from the first measurement (performed in a measurement room as described in Appendix B.1) can be seen in Figures 12 and 13 as scatter plots, showing the TOA along the x- and the DOA along the y-axis.

Each scatter plot in Figures 12a, 12c and 12e shows the DOA and TOA estimation results for one of the three available microphone. The direct sound of each source $j \in \{A, B, C\}$ arrives at a certain microphone i_0 first, which is closest to that source. This arrival is used as a reference for computing the direct sound TDOAs $\Delta t_{i,j}^{(i_0)}$ at all other microphones. The TDOA at the reference microphone i_0 for the respective source computes to zero.

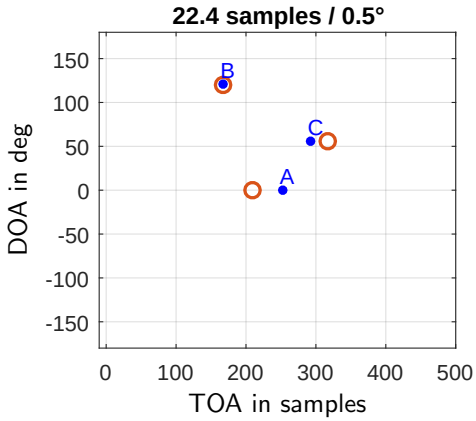
The correct TOA-DOA pairs are computed from the measured positions shown in Figure 54 and are indicated as blue circles marked with the corresponding source letter $\{A, B, C\}$, the estimated pairs are indicated by a red circle. The results are compared using the mean-absolute-error at each microphone for all sources j according to the formula

$$\epsilon_x = \frac{1}{N_j} \sum_{j=1}^{N_x} |x_{j,est} - x_{j,true}| \tag{2.4.1}$$

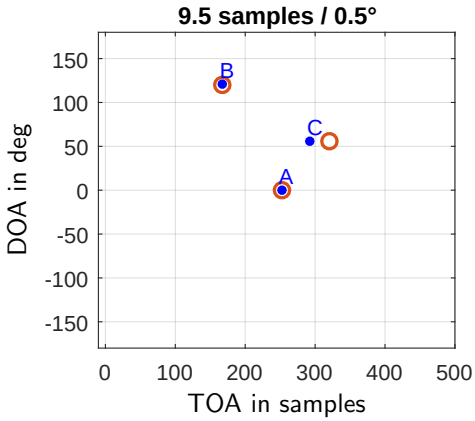
where $x_{j,true}$ is used as substitute for the true TOA and DOA parameters and $x_{j,est}$ for the respective estimated parameter values. The results for the errors are summarized in Table 2. The sample rate was $f_s = 44100\ Hz$ and the speed of sound was assumed to be $c = 340\ \frac{m}{s}$.

Table 2: Comparison of the TDOA and DOA error for the broadband (bb), low pass (lp) and high pass (hp) clap types.

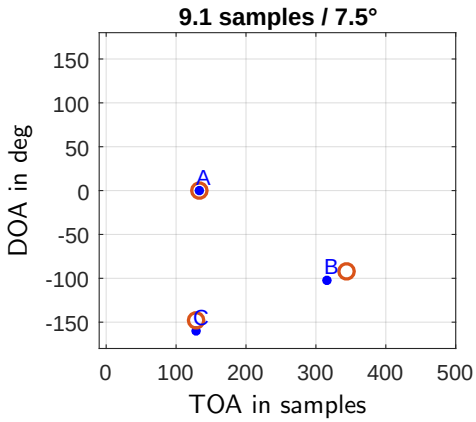
mic.	*clap type*	ϵ_{TOA} in samples	ϵ_{DOA} in deg
1	bb	22.4	0.5
2	bb	9.1	7.5
3	bb	33	11.1
1	lp	9.5	0.5
2	lp	3.8	3.5
3	lp	9	7.1
1	hp	7.7	0.5
2	hp	10.5	3.5
3	hp	7.7	8.4



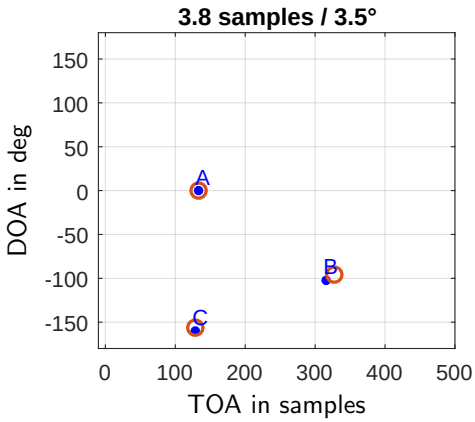
(a) $i = 1$, bb



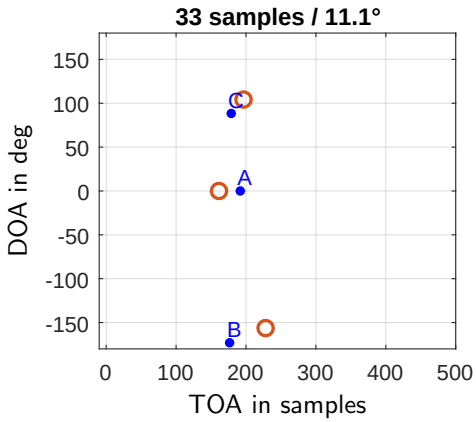
(b) $i = 1$, lp



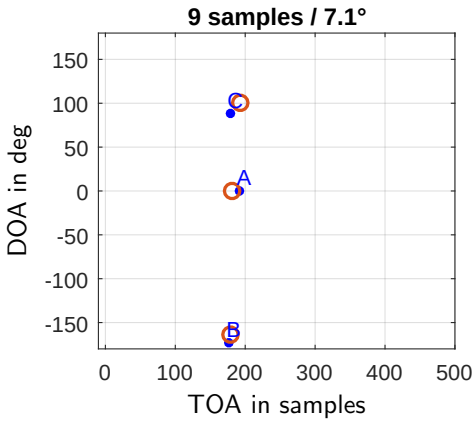
(c) $i = 2$, bb



(d) $i = 2$, lp

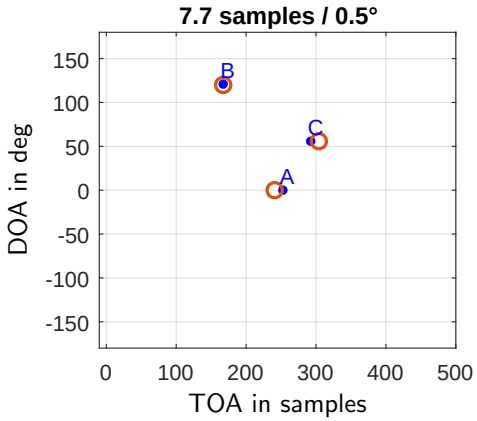


(e) $i = 3$, bb

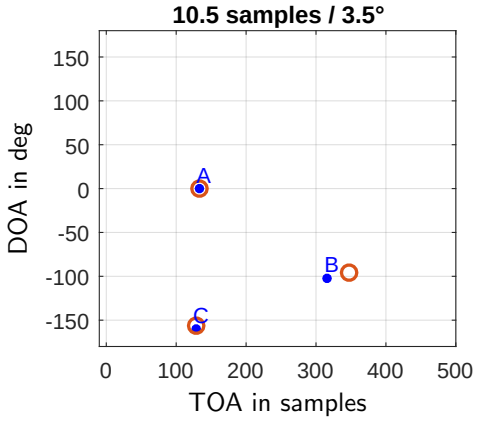


(f) $i = 3$, lp

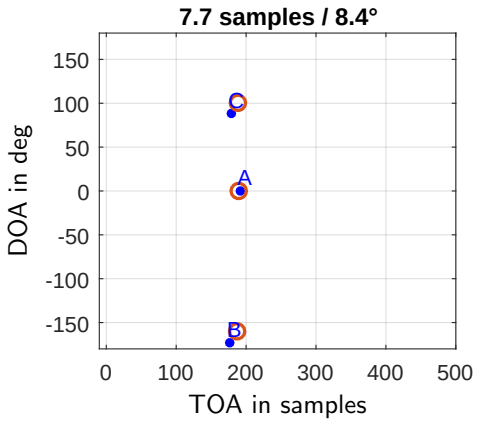
Figure 12: Results of DOA and TOA estimates using the data form the first measurement (Appendix B.1) for the broadband and low pass type clap. The corresponding errors are shown above the plots as numerical values ϵ_{TOA} *samples* / ϵ_{DOA} °. The measured values are marked as ●, the estimated ones by ○.



(a) $i = 1$, hp



(b) $i = 2$ hp



(c) $i = 3$, hp

Figure 13: Results of DOA and TOA estimates using the data form the first measurement (Appendix B.1) for the high pass type clap. The corresponding errors are shown above the plots as numerical values ϵ_{TOA} *samples* / ϵ_{DOA} °. The measured values are marked as ●, the estimated ones by ○.

3 Scene Reconstruction

In this chapter, the estimated parameters from Chapter 2 are used to construct a model of the acoustic scene. The resulting model will contain the microphones, calibration source positions and the boundaries of the room, i.e. the four surrounding walls as well as floor and ceiling in the case of the second measurement, and the single reflector for the first measurement. The scene reconstruction is independent of the implementation of the parameter estimation, as long as DOA and TOA pairs (of direct sound and an arbitrary number of reflections) are available. Optional parameters are weights used for sorting the TOA-DOA pairs of the reflections according to how accurate the estimated reflection parameters are expected to be.

After initially giving a short overview of literature the self-calibration and room inference problem in Section 3.1, this chapter describes the proposed self-calibration of the microphone arrays (Section 3.2) and the room inference, which is separated into the localization of reflecting surfaces (Section 3.3) and the estimation of the room height (Section 3.4).

3.1 Literature Review

When attempting to perform geometry inference, knowledge of the positions of the microphones or microphone arrays used is integral. The process of trying to recover these positions using only signals recorded by the microphones of a number of calibration sources (instead of performing manual measurements), is usually called self-calibration or geometry-calibration in literature. For this task a number of algorithms exist, which usually rely on timing information of the signals arriving at each microphone or microphone array, i.e. time differences between recorded signals. More seldom used information is the direction from which the source signals arrive. These time and direction parameters are usually referred to as *time of arrival* (TOA) and *direction of arrival* (DOA) respectively.

Examples for self-calibration can be found in the works by Gaubitch et al. in [GKH13] where the self-calibration is performed using the source TOAs at microphone signals or by Pollefeys and Nister in [PN08] where the time of departure of sound at a calibration source is computed from measured TDOAs which are then used for the self-calibration. A closed form solution for the self-calibration can be found when

assuming that a source and a microphone are co-located as shown by Crocco et al. in [CDM12]. Valente et al. [VTA+10] propose a method for self-calibration (therein called geometric calibration) of two unsynchronized arrays by performing source localization of a calibration source and then fusing these estimates together. Schmalenströer et al. [SJHU+11] introduce an algorithm using DOA and TDOA information for the self-calibration. Another method that makes use of a special calibration source (a loudspeaker with a microphone attached above its centre) and three reference microphones (with known/measured coordinates) for self-calibration is described by Khanal in [KSS13].

A source localization method using unsynchronized arrays is described by Hack in [Hac15] alongside an examination of the DOA estimation capabilities of the used first order microphone arrays in three dimensions and an examination of the DOA estimation errors introduced by the array geometry.

The next logical step after self-calibration and source localization is performing acoustic geometry inference or room inference, where the locations of reflecting surfaces are estimated, either by using known (e.g. measured) microphone array locations or after performing self-calibration. In case of geometry inference, most algorithms either rely on TOA or DOA data acquired from discrete microphones or microphone arrays, combined by either localizing real as well as mirror sources (a setup including mirror sources is shown in Figure 14) and then finding the corresponding reflector between the real and the mirror source, or by directly localizing the reflection points. A thorough investigation of time based source localization and room inference is given by Filos [Fil13] who uses the TDOAs obtained from the TOAs of direct sound and reflections at a microphone array to localize reflective surfaces via a common tangent of multiple ellipses representing all possible locations of a reflector fulfilling the found TDOAs. As calibration source signals finger snaps or loudspeakers emitting MLS signals to measure impulse responses are used. Another TOA based approach is shown by Tervo et al. using either continuous signals (e.g. speech or music) and solely time differences in [TK10] or by the use of measured RIRs to localize reflection points via DOA and TOA estimates in [TKL11]. An approach for the joint use of DOA and TDOA data for reflector localization is proposed by Sun et al. in [HEK11] using spherical microphone arrays to find the DOA of a direct sound source and then using the extracted direct signal for localizing reflections of first order using the reflection DOAs and TDOAs (the latter of which are obtained by correlation).

Another example in the field of room inference (there termed room reconstruction) is described by Dokmanic et al. in [DDV16], where the concept of *simultaneous localization and mapping* (SLAM) is translated to the acoustic domain. Therein

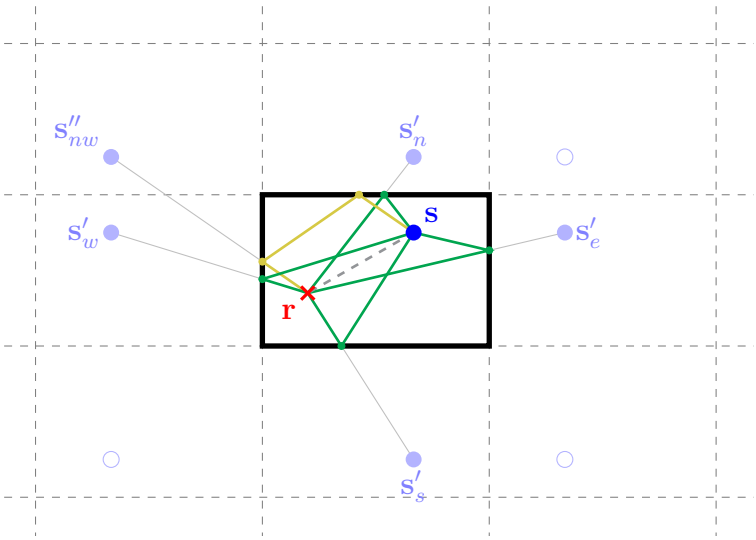


Figure 14: Room model showing a microphone **r**, a source **s** and all first order mirror sources $s'_{\star}$, one second order source $s''_{\star\star}$ as well as direct and reflection paths. The walls are termed north, east, south and west (n,e,s,w).

Dokmanic et al. show time based room inference under the assumption of known array geometry by localizing mirror sources, self-calibration of a microphone array (shown to be solvable up to an arbitrary rotation without any prior information) using the direct sound and reflections, as well as the combination of the results of the latter for tracking a moving robot equipped with a microphone which records the rooms responses to excitation by a loudspeaker. SLAM was initially introduced in the field of robotics (using radio transmissions) for localizing the boundaries of a room as well as tracking the movements of a robot within that same room (usually with no or little prior knowledge). Descriptions of initial algorithms can be found in Smith et al. [SSC90] and Leonard and Durrant-Whyte [LDW91, LDWC92] utilizing different sonar setups and usually Kalman filtering.

A framework for performing recordings with microphone arrays that can be distributed wirelessly within an acoustic scene was developed at the *Institute of Electronic Music and Acoustics* (IEM) in form of the WiLMA (*wireless large-scale microphone array*) project by Schörkhuber et al. as described in [SZZ14]. The microphone arrays are connected to a central unit by means of wireless modules (a module is shown in Figure 15). Each wireless module processes the signal of one B-format microphone with a sampling frequency of $48kHz$ and a synchronization error between the modules in the range of 1 *sample* $\approx 20\mu s$. Additionally, they provide the power for the microphone arrays and handle recording duties and allow simple signal processing tasks.

Figure 15: The sensor modules of the WiLMA on the left, together with the `Oktava Ambient 4D` B-Format microphones used, taken from [WIL].

A comparable system is implemented and examined by Cobos et al. [CPSFC+14], examining algorithms for a *wireless acoustic sensor network* and performing source localization from arbitrary signals of discrete microphones and known sensor locations. The source localization is performed using TDEs obtained by a cumulative sum based onset detection, with the applicability in low-cost networks in mind.

Corresponding to the different algorithms for TOA and DOA estimation also different signal types are used, depending on the capabilities and requirements of the respective algorithms. Popular are the use of a *room impulse response* (RIR) as described by Tervo and Politis [TP15] and Filos [Fil13] or a continuous signal (usually speech or music) as described by Tervo and Korhonen [TK10]. As measuring of RIRs can be tedious, for example using sweeps with a measuring setup consisting of at least one loudspeaker and a microphone needed, claps or other impulse-like sounds as for example in [Fil13] that can be produced mechanically are often used to produce signals comparable to impulse responses.

Another interesting scene parameter is the room volume, which is examined using either RIRs [SZR10] or recordings of reverberant speech [SRZ10,SZR13] for classifying rooms using *mel frequency cepstral coefficients*, features popularly used in speech processing.

An example of an acoustic scene with multiple moving sources and microphone locations, is given in Figure 16, although with no reflectors present. Possible reflectors could be walls of houses of similar things.

After this review the most important terms are summarized for clarification: *microphone self calibration* describes the act of finding the positions of a number

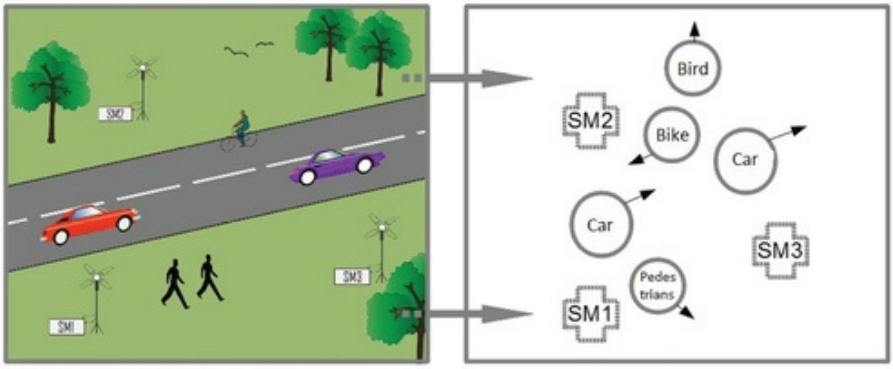


Figure 16: An example of an arbitrary acoustic scene on the left, with a corresponding scene model on the right, taken from [WIL], in this case without reflecting surfaces.

microphones and calibration sources with the main focus on the microphones since the sources are only needed specifically for the calibration. *Room inference* deals with finding the geometry of the surroundings of a microphone array, not necessarily including the self-calibration. An *acoustic scene* is usually made up of one or more moving sources (see Figure 16) while an *acoustic system* comprises a microphone array or a number of microphones or microphone arrays and a geometry of sort producing reflections.

3.2 Self-calibration

The proposed self-calibration algorithm simultaneously estimates the positions of microphones and calibration sources (in two dimensions) up to an arbitrary rotation. It only needs the TOAs and the corresponding DOAs of the direct sounds to compute *time* and *direction differences of arrival* (TDOAs and DDOAs). The microphone arrays therefore need to be synchronised as presumed throughout this work.

The microphone and source positions are described as complex numbers to describe points in the 2D where the x coordinate is given by the real and the y component by the imaginary[5] part, according to Equations 1.1.7, 1.1.8 and 1.1.9. Due to the fact that the centre of the coordinate system can be chosen arbitrarily, it makes sense to position an arbitrary source in the origin (*first assumption*). Since the orientation of each microphone array, described by the direction of the DOA with $\upsilon = 0°$, is usually neither known nor easily measurable all microphones are orientated towards

[5]the letter i will be used as microphone index and as the imaginary unit $i = \sqrt{-1}$, which will be clear by context

Assumption 1:	an arbitrary source is chosen as the centre of the coordinate system (reference source j_0)
Assumption 2:	all microphones are assumed to be oriented (0° direction) towards the reference source
Assumption 3:	an arbitrary microphone is fixed to the real axis (reference microphone i_{ref})

Figure 17: Assumptions for the self-localization task.

the source chosen as origin (*second assumption*). Similar to the TDOAs between arrivals of acoustic events, the DDOA $\Delta\upsilon_{i,j}^{(j_0)}$ is used describing the angle between the reference DOA (direct sound from source j_0) and any other direct sound DOA of source j at microphone i, according to

$$\Delta\upsilon_{i,j}^{(j_0)} = \upsilon_{i,j} - \upsilon_{i,j_0}. \tag{3.2.1}$$

To simplify the notation and because j_0 is the same for all microphones the index can be dropped for readability resulting in $\Delta\upsilon_{i,j}^{(j_0)} \equiv \Delta\upsilon_{i,j}$

The *third assumption* is that the a microphone i_{ref} is located on the real axis, resulting in real coordinates $\bar{z} = r$. All three assumptions are summarized in Figure 17.

Throughout this work the reference source and microphone are chosen as the first one respectively, resulting in $j_0 = 1 = A$ and $i_{\text{ref}} = 1$.

From the parameters estimated at each microphone a phasor system containing calibration source estimates and said microphone in a local coordinate system can be constructed, shown for an exemplary setup in Figure 19. The basic thought of the proposed self-calibration algorithm is to find the alignment of all phasor systems (one for each microphone) such that the respective estimated source positions from all phasor systems align perfectly (in case of known TDOAs and DDOAs), or as good as possible (when using estimates which are usually not ideal). The estimated parameters are the direct sound DDOAs and TDOAs $\Delta\upsilon_{i,j}$ and $\Delta t_{i,j}^{(i_0)}$, the yet unknown parameters are the rotation angles $\breve{\varphi}_i$ needed to align the phasor systems of microphones i and, since only the TDOAs between the direct sounds of sources can be estimated, the *time of flight* (TOF) that describes the distance from a source j to the closest microphone described by $\delta_{i_0,j}$. Section 3.2.1 describes these parameters $\breve{\varphi}_i$ and $\delta_{i_0,j}$ in detail as well as a possible way to find them by minimizing a suitable cost function.

3.2.1 Proposed Self-calibration Algorithm

The estimated positions of microphones and sources can be computed from the parameter set minimizing a multidimensional cost function $J(\mathbf{\Theta}, \mathbf{R})$ with

$$\mathbf{\Theta} = \begin{pmatrix} 0 & \breve{\varphi}_2 & \breve{\varphi}_3 & \cdots & \breve{\varphi}_{N_i} \end{pmatrix} \tag{3.2.2}$$

and

$$\mathbf{R} = \begin{pmatrix} \delta_{i_0,1} & \delta_{i_0,2} & \delta_{i_0,3} & \cdots & \delta_{i_0,N_j} \end{pmatrix}. \tag{3.2.3}$$

$\mathbf{\Theta}$ contains the rotation angles $\breve{\varphi}_i$ of the microphone phasor systems with $\breve{\varphi}_1 = 0$ (*Assumption 3*) which reduces the microphone rotation parameters $\mathbf{\Theta}$ by one. The angles stored in $\mathbf{\Theta}$ are needed because the orientation of the microphones is unknown, thus each array on its own only knows the calibration source directions in its own coordinate system. Another interpretation of $\breve{\varphi}_i$ is that it describes the rotations that are needed to combines each local coordinate system in which sources are estimated relative to one microphone array into a mutual coordinate system where all estimates of a source j are as close together as possible.

The TOFs stored in $\mathbf{R}$ are needed because the estimated TDOAs $\Delta t_{i,j}^{(i_0)}$ only define the distance between a microphone i and source j up to said unknown TOF $\delta_{i_0,j}$ of the source j to the closest microphone i_0 (with $\Delta t_{i_0,j}^{(i_0)} \stackrel{!}{=} 0$). The distance between each source j and microphone i can thus be expressed as $\Delta t_{i,j}^{(i_0)} + \delta_{i_0,j}$. It is important to notice that in contrast to j_0 which was the same for all microphones, i_0 can vary from source to source and is therefore a function of j according to

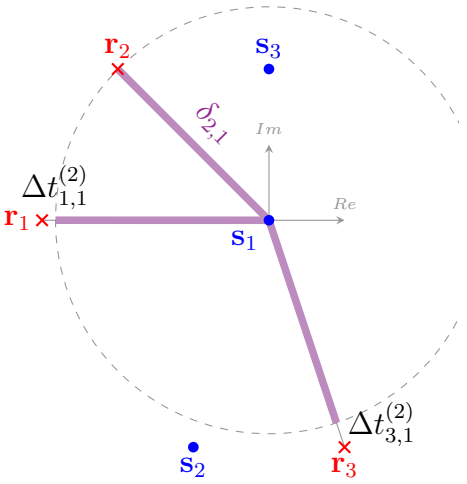
$$i_0(j) = \underset{i}{\operatorname{argmin}} \{t_{i,j}\} \tag{3.2.4}$$

which gives the index of the microphone at which the earliest TOA $t_{i,j}$ of the direct sound emitted by source j was detected. $i_0(j)$ indicates which microphone is closest to source j. This was already described in Section 1.1 but is repeated here for clarification. Where the entries of the parameter vectors can be found in a geometric sense is shown in Figure 18 for $\mathbf{R}$ and in Figures 19 and 20b for $\mathbf{\Theta}$.

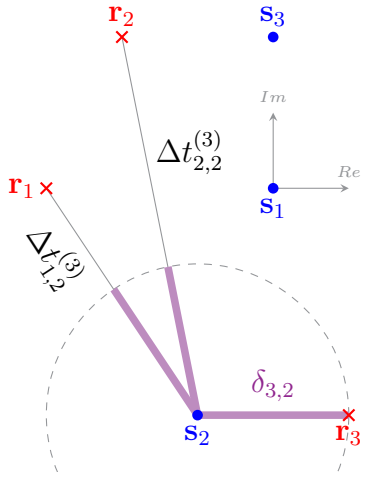
Finding the optimal overlap of all the local coordinate systems of all microphones using the distances in $\mathbf{R}$ and angles in $\mathbf{\Theta}$ is what is achieved by minimizing the cost function that is described below.

The position z_i of a microphone i on the real axis in its own local coordinate system can be written as

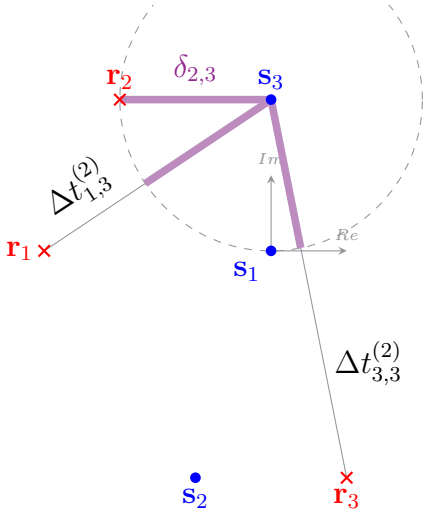
$$z_i(\delta_{i_0,j_0}) = -(\Delta t_{i,j_0}^{(i_0)} + \delta_{i_0,j_0}), \tag{3.2.5}$$



(a) $\mathbf{s}_1$: $\delta_{2,1}$ with $i_0(j=1)=2$

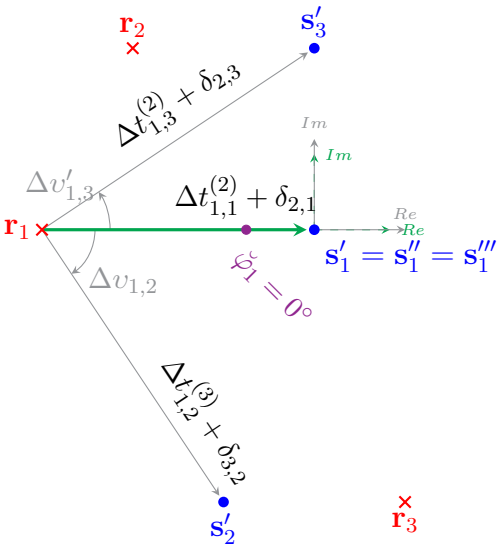


(b) $\mathbf{s}_2$: $\delta_{3,2}$ with $i_0(2)=3$

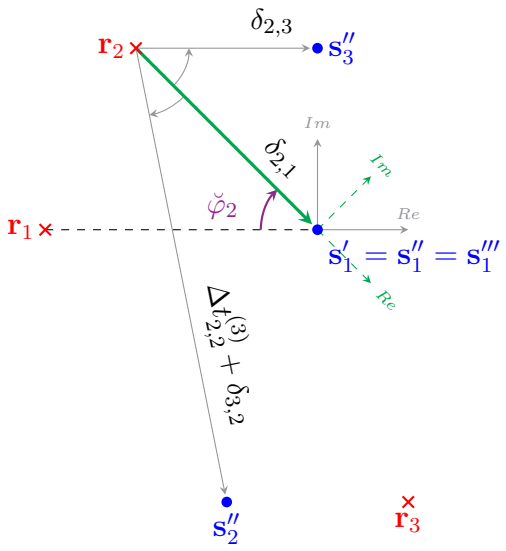


(c) $\mathbf{s}_3$: $\delta_{2,3}$ with $i_0(3)=2$

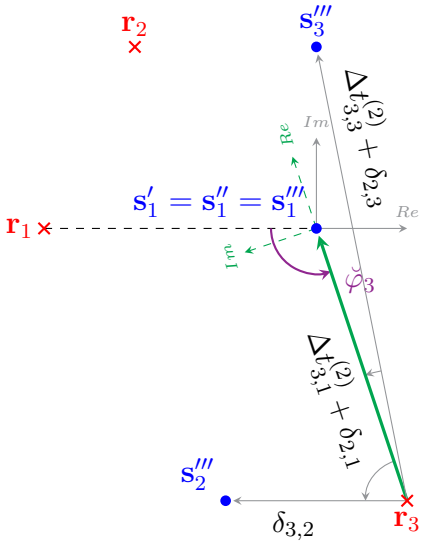
Figure 18: Self-calibration parameters $\delta_{i_0,j}$ stored in $\mathbf{R}$ that are source dependent, shown for an exemplary setup. The reference microphone for TDOA computation for each source j is the one on the circle centred around the respective source. The coordinate axes of mutual coordinate system are indicated in grey with the origin ate the reference source $\mathbf{s}_1$.



(a) $\mathbf{r}_1$: $\breve{\varphi}_1 = 0°$



(b) $\mathbf{r}_2$: $\breve{\varphi}_2$



(c) $\mathbf{r}_3$: $\breve{\varphi}_3$

Figure 19: Self-calibration parameters $\breve{\varphi}_i$ stored in $\boldsymbol{\Theta}$ that are microphone dependent. The source dependent parameters $\delta_{i_0,j}$ from $\mathbf{R}$ are indicated in the microphone-source distances as well. The mutual coordinates are indicated by the grey axes, the local coordinates for each microphone by the dashed green axes.

where j_0 is the index of the reference source and δ_{i_0,j_0} the TOF of reference source j_0 to its closest microphone i_0. From each microphone position z_i the position $\bar{z}_{i,j}$ of source j can be found via the estimated parameters (direct sound TDOAs and DDOAs) in the local coordinate system of the respective microphone i as

$$\bar{z}_{i,j}(\mathbf{R}) = z_i(\delta_{i_0,j_0}) + (\Delta t_{i,j}^{(i_0)} + \delta_{i_0,j}) \cdot e^{i \cdot \Delta \upsilon_{i,j}}, \tag{3.2.6}$$

where $\Delta \upsilon_j$ is the DDOA between the DOA of the direct sound of the chosen reference source j_0 and a source j. From the microphone position $z_i(\delta_{i_0,j_0})$ all sources of a phasor system can be computed in the local coordinate system of microphone i.

Concluding that all microphones were able to detect all direct sounds, every source location $z_{i,j}$ can be described from each of the N_i microphones up to the correct rotation $\breve{\varphi}_i$ (see Figure 19) of the local microphone coordinates as

$$\begin{aligned} z_{i,j}(\boldsymbol{\Theta}, \mathbf{R}) &= \left(z_i(\delta_{i_0,j_0}) + (\Delta t_{i,j}^{(i_0)} + \delta_{i_0,j}) \cdot e^{i \cdot \Delta \upsilon_{i,j}} \right) \cdot e^{i \cdot \breve{\varphi}_i} \\ &= \bar{z}_{i,j}(\mathbf{R}) \cdot e^{i \cdot \breve{\varphi}_i}. \end{aligned} \tag{3.2.7}$$

The complex number $z_{i,j}$ then describes the location of source j relative to microphone i. When the correct microphone rotation angles $\breve{\varphi}_i$ (and TOFs) are known and all estimates are perfect, all source position estimates align as shown in Figure 20b and all microphones are rotated to their actual positions. The orientation of each local phasor systems of the example system is indicated as well, centred around the respective microphone for clarification.

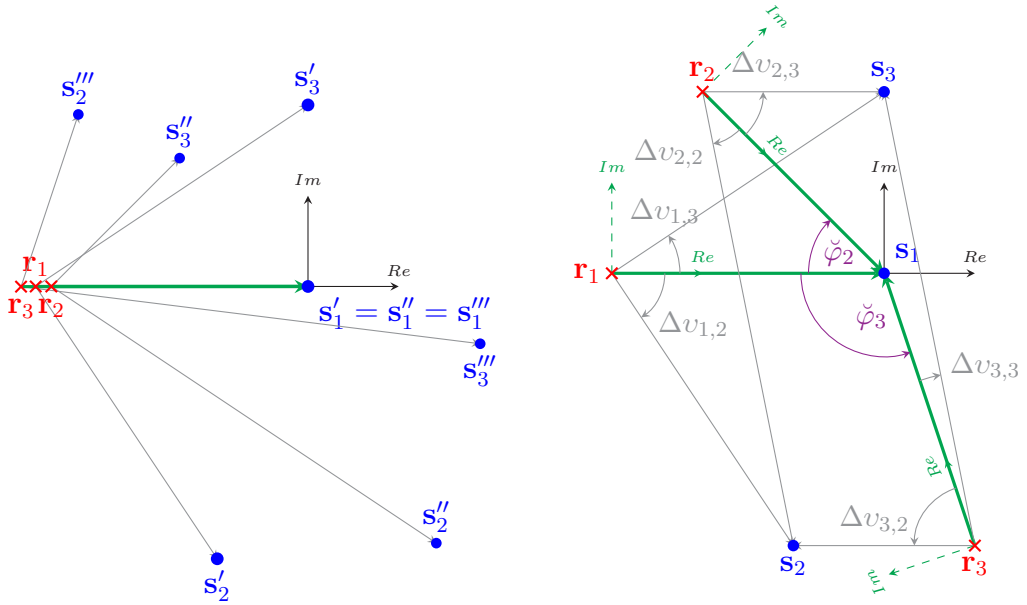
The positions $z_{i,j}(\boldsymbol{\Theta}, \mathbf{R})$ of all source estimates can be combined in the complex position matrix

$$\mathbf{P}(\boldsymbol{\Theta}, \mathbf{R}) = \Big\{ z_{i,j}(\boldsymbol{\Theta}, \mathbf{R}) \Big\}_{i,j}, \tag{3.2.8}$$

with $j \in \{1, \ldots, N_j\}$ denoting the row index and $i \in \{1, \ldots, N_i\}$ denoting the column index (resulting in a $(N_i \times N_j - 1)$-matrix) with the estimated position source j with respect to microphone i at the position (i, j) in the matrix. Since this yields N_i position estimates per source, averaging over all rows results in the final source position estimate.

The cost function to minimize $J(\boldsymbol{\Theta}, \mathbf{R})$ can thus be written as

$$J(\boldsymbol{\Theta}, \mathbf{R}) = \sum_{j=1}^{N_j - 1} \sum_{w=1}^{N_i} \sum_{\substack{i=1 \\ i \neq w}}^{N_i} \Big| z_{w,j}(\boldsymbol{\Theta}, \mathbf{R}) - z_{i,j}(\boldsymbol{\Theta}, \mathbf{R}) \Big|, \tag{3.2.9}$$



(a) starting assumption **(b)** optimal positioning

Figure 20: Phasor diagram of a hypothetical setup with three microphones $\mathbf{r}_i$ and three sources $\mathbf{s}_j$, with known microphone rotation angles $\breve{\varphi}_i$ (Figure 20b) and of the starting assumption with $\breve{\varphi}_i = 0 \;\forall\; i \in \{1, 2, 3\}$ (Figure 20a). $\mathbf{s}_j'$ represents source j estimated from microphone 1 and so forth.

which describes the sum of the differences between the source estimates of each source form different microphones.

A hypothetical setup for three microphones and three calibration sources is depicted in Figure 20. Figure 20a shows the three microphone phasor systems aligned as defined in *Assumption 1* with all microphones initially oriented towards the origin as described by *Assumption 2*. Figure 20b shows the correct setup after performing minimization on the cost function $J(\mathbf{\Theta}, \mathbf{R})$.

Finding the optimal microphone rotation angles $\breve{\varphi}_i$

To find the correct rotation angles $\breve{\varphi}_i$ of the microphones as displayed in Figure 20b (initially assuming the TOFs in $\mathbf{R}$ are known for the sake of explanation), all microphone rotation angles are set to $\breve{\varphi}_i = 0 \;\forall\; i$ such that all the microphones are located somewhere on the negative real axis (see Figure 20a). Each microphone $i \in \{1, \ldots, N\} \setminus i_{\text{ref}}$ is then rotated from 0 to 2π until minimum distances between this and all other microphones respective source estimates is achieved. The resulting rotation angles $\breve{\varphi}_i$ of each iteration are stored in $\mathbf{\Theta}^{(a)}$, with a indexing the iteration.

One iteration consists of rotating all microphones $N_i - 1$ once in consecutive order and saving the parameters achieving minimum distance between all source estimates. Iterations are continued as long as $J(\mathbf{\Theta}^{(a)}, \mathbf{R}) < J(\mathbf{\Theta}^{(a-1)}, \mathbf{R})$ (i.e. as long as the cost function decreases) or until a maximum number of iterations $a_{\max}$[6] was performed. The greedy stopping criterion of constantly reducing the cost function was found to be applicable because for a fixed angle resolution of the rotation angles $\breve{\varphi}_i$ the cost function saturated at a value slightly higher than the actual minimum.

Finding the optimal TOFs $\delta_{i_0,j}$

For the real case of unknown TOFs, optimizing the microphone rotations $\breve{\varphi}_i$ alone is obviously not sufficient for locating the sources and microphones. A solution can be found quite easily by alternatively minimizing the cost function $J(\mathbf{\Theta}, \mathbf{R})$ with respect to $\mathbf{\Theta}$ and $\mathbf{R}$, starting with

$$\hat{\mathbf{\Theta}}^{(1)} = \underset{\mathbf{\Theta}}{\operatorname{argmin}} \left\{ J\left(\mathbf{\Theta}, \mathbf{R}^{(0)}\right) \right\}, \quad i.w.\ \mathbf{\Theta} = \mathbf{0} \tag{3.2.10}$$

where *i.w.* stands for *initialized with*, $\mathbf{0}$ is a zero-vector of size $(1 \times N_i)$ and $\mathbf{R}^{(0)}$ contains the initialization for the TOFs. These are set to an identical (but arbitrary) positive value to ensure that no microphone is co-located with the chosen reference source j_0. The minimum of the cost function yields the estimate for the microphone rotations $\hat{\mathbf{\Theta}}^{(1)}$ after the first iteration (obtained by consecutively rotating all microphones as described above), which is then used to find the optimum value for $\hat{\mathbf{R}}^{(1)}$ according to

$$\hat{\mathbf{R}}^{(1)} = \underset{\mathbf{R}}{\operatorname{argmin}} \left\{ J\left(\hat{\mathbf{\Theta}}^{(1)}, \mathbf{R}\right) \right\}. \quad i.w.\ \mathbf{R} = \mathbf{R}^{(0)} \tag{3.2.11}$$

By initializing the next iteration $p+1$ with the optimal values from the last one and continuing in an alternating fashion we get

$$\hat{\mathbf{\Theta}}^{(p+1)} = \underset{\mathbf{\Theta}}{\operatorname{argmin}} \left\{ J\left(\mathbf{\Theta}, \hat{\mathbf{R}}^{(p)}\right) \right\} \quad i.w.\ \mathbf{\Theta} = \hat{\mathbf{\Theta}}^{(p)}, \tag{3.2.12}$$

[6] The algorithm always converged before reaching the maximum number of iterations usually chosen with $a_{\max} = 30$.

and

$$\hat{\mathbf{R}}^{(p+1)} = \underset{\mathbf{R}}{\operatorname{argmin}} \left\{ J\left(\hat{\boldsymbol{\Theta}}^{(p+1)}, \mathbf{R}\right) \right\} \quad (3.2.13)$$
$$i.w.\ \mathbf{R} = \hat{\mathbf{R}}^{(p)}.$$

The minimization with respect to $\mathbf{R}$ was performed using `CVX`, a package for specifying and solving convex programs [GB14, GB08].

As stopping criterion for the alternating optimization the number of times a that all $N_i - 1$ microphones are rotated during one iteration p was used.

3.3 Room Inference

A popular approach on the room inference problem is to localize reflective surfaces (i.e. walls) by localizing mirror sources, which proved difficult because the microphone arrays usually only detected a single strong reflection and only few microphones detected the reflection of the same wall (i.e. the same mirror source). Therefore another approach found in the literature is used in this section, namely finding reflection points satisfying the found TDOA and DDOA pairs assigned to reflections, followed by a stage that searches for similarities and characteristic features in the found points. The estimation of reflection points is similar to time based reflector localization algorithms that work on TDOAs alone, with the advantage that due to the known DDOA of the reflection the reflection point can be computed directly using trigonometric identities.

As already mentioned, the room inference stage needs the location map of the microphones and sources in addition to the reflection DDOA and TDOA estimate pairs (for an arbitrary number of microphones) to be able to correctly combine its results with the source and microphone positions.

The computation of reflection point candidates is described in Sections 3.3.1 and 3.3.2 followed by a description of four algorithms that try to extract the room shape from the computed reflection points. Two of these explicitly include the assumption of a rectangular room (Sections 3.3.5 and 3.3.6) and two try to detect linear reflectors of arbitrary orientation (Sections 3.3.3 and 3.3.4).

3.3.1 Reflection Point Computation

With the microphones and sources localized (e.g. using the algorithm described in Section 3.2) and the TDOA and DDOA pairs of reflections available, the reflection

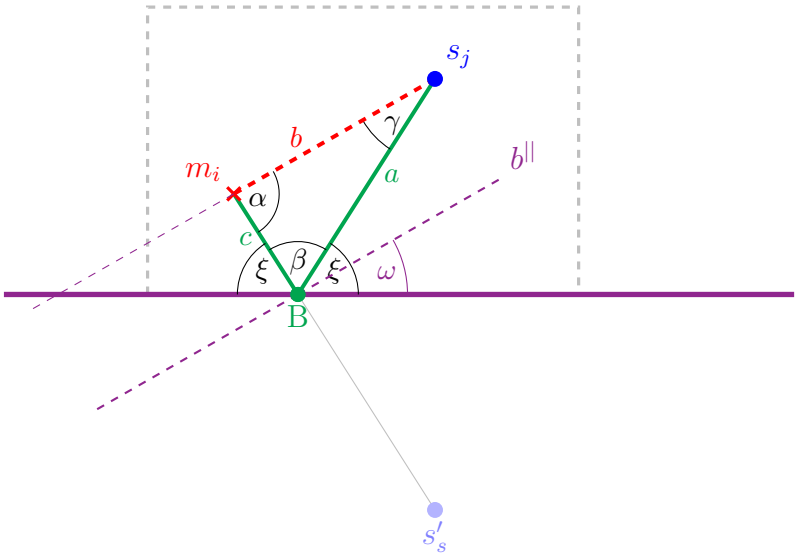


Figure 21: Triangle used to compute a single reflection point **B**, where α is the reflection DDOA $\Delta\underline{\upsilon}_{i,j}^{(\ell)}$ and ω represents the angular orientation of the reflector. The entire room is shown in **gray**, the reflector to find in **violet**, the reflection path in **green (a,c)** and the direct path in **red (b)**.

points and the orientation of a linear reflecting surface can be computed using simple trigonometric equations. It should be noted that only first order reflections can be computed directly from the estimated parameters[7]. An example for finding a single reflection point can be seen in Figure 21, where α represents the reflection DDOA $\Delta\underline{\upsilon}_{i,j}$ and ω the orientation angle of the reflector. The hypothetical room is depicted by a dashed rectangle, the path of the direct sound is indicated as b, the path of the reflection with a and c and the reflection point by B.

The sides of the triangle represent the distance the direct sound covers from source j to the microphone i as $b = ||\mathbf{r}_i - \mathbf{s}_j||_2$ and the distance the reflection travels as

$$a + c = b + \Delta t_{i,j}^{(i_0)}, \tag{3.3.1}$$

with c being the quantity of interest, which can be computed directly as

$$c = \frac{2b \cdot \Delta t_{i,j}^{(i_0)} + (\Delta t_{i,j}^{(i_0)})^2}{2(b + \Delta t_{i,j}^{(i_0)}) - 2b\cos\alpha}, \tag{3.3.2}$$

[7]Higher order reflection points that are wrongly assumed to be of first order will always be farther away then a first order source detected from the same direction and thus pose no real problem.

using the law of cosines

$$a^2 = b^2 + c^2 - 2bc\cos\alpha \tag{3.3.3}$$

with Equation 3.3.1 inserted for the unknown a.

With the angles in triangles usually defined positive $\alpha = |\Delta\underline{\upsilon}_{i,j}^{(\ell)}|$ is used, with the orientation ω of the reflector defined relative to a line parallel to the DOA of the detected direct sound of source j(denoted by $b^{\parallel}$ in Figure 21). This results in

$$\omega = \operatorname{sgn}\left(\Delta\underline{\upsilon}_{i,j}^{(\ell)}\right) \cdot \left(\alpha - \xi\right), \tag{3.3.4}$$

with the correct sign for all reflection DDOAs. $\operatorname{sgn}(\cdot)$ is the signum function defined as

$$\operatorname{sgn}(x) = \begin{cases} +1, & x > 0 \\ 0, & x = 0 \\ -1, & x > 0. \end{cases} \tag{3.3.5}$$

These equations are then used to compute all reflection points and corresponding angles for the found reflection TDOA and DDOA pairs in the same coordinate system defined by the reference source and microphone.

3.3.2 Reflection Point Separation

The reflection points found in Section 3.3.1 are then separated into two groups: those with TDOA-DDOA pairs fulfilling *both* constraints

$$|\Delta\underline{\upsilon}_{i,j}^{(\ell)}| < D^{\circ} \tag{3.3.6}$$

$$\Delta t_{i,j}^{(\ell)} < \frac{2 \cdot ||\mathbf{r}_i - \mathbf{s}_j||_2}{c}, \tag{3.3.7}$$

and those *not* fulfilling both constraints.

The constraints represent a minimum DDOA and TDOA between reflections and the direct sound which will be needed for reflections to be caused by walls and not by floor or ceiling. The allowed angle difference was usually chosen with $D = 10^{\circ}$. A simple example makes the problem with said these constraints obvious.

Constraint Example: Assume that a microphone to source distance of 1 m was found and the (unknown) distances of the microphone to floor and ceiling are both 1.5 m, the resulting distance a floor or ceiling reflection travels would result in $d_r = 2 \cdot \sqrt{0.5^2 + 1.5^2} = 2 \cdot 1.5811 = 3.1662$ m compared to a distance of 1 m which already would not be excluded by above constraints. Assuming an actual microphone source distance of 1.5 m, the constraints already work: $d_r = 2 \cdot \sqrt{0.75^2 + 1.5^2} = 2 \cdot 1.6771 = 3.3541$ m compared to a distance of 1.5 m.

Choosing constraints on the TDOAs is obviously dependent on the room height, though having accurate elevation DOA estimates at hand would render these constraints unnecessary.

The TDOAs of the points that are fulfilling said constraints are stored in the set $\mathcal{P}_{c,f}$. They are assumed to correspond to floor and ceiling and are examined in Section 3.4 using them to acquire estimates of the room height.

To acquire a working set for performing room inference, all the reflection points *not* fulfilling the constraints given by Equations 3.3.6 and 3.3.7 are stored as the set of points $\mathcal{P}$. Points inside the convex hull (see Figure 22) of all estimated microphone and calibration source estimates are removed from the point set $\mathcal{P}$ and stored in a new set $\bar{\mathcal{P}}$. This set now contains possible reflection points that lie in between estimated source and microphone positions (created by e.g. tables, ...). The points in $\bar{\mathcal{P}}$ could be examined later on to find possible reflecting surfaces that are scattered throughout the scene, although this will not be part of this work.

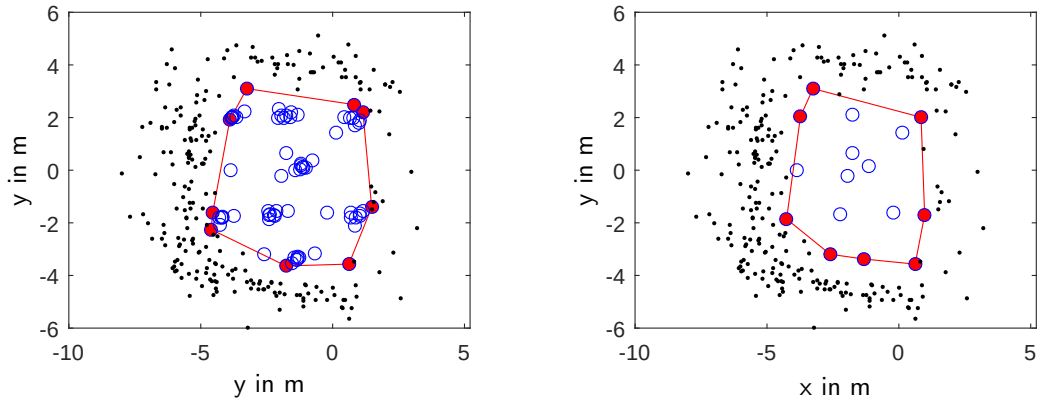
The points remaining in the set $\mathcal{P}$ are used for the following room inference algorithms. Each reflection point from the $\mathcal{P}$ is furthermore assigned a weight corresponding to the density of the DOA estimates (indicating the similarity between the DOA estimates of all frequencies) for the corresponding reflection. This weight indicates sharpness of the DOA estimate of the reflection and is computed using Equation 2.3.5, which is

$$\rho(k,m) = \sum_{\kappa} \exp\left(-\frac{d\big(\upsilon(k,m),\upsilon(\kappa,m)\big)^2}{d_c^2}\right) \tag{3.3.8}$$

and then taking the maximum thereof

$$\rho_{max}(m) = \max_{k} \rho(k,m), \tag{3.3.9}$$

with the corresponding arrival time of the reflections inserted for m to get the weight $w_{\mathcal{P}}$ of the reflection point from the set $\mathcal{P}$.



(a) all estimates **(b)** averaged estimates

Figure 22: Polygons $\mathcal{G}$, used to find points for $\bar{\mathcal{P}}$ located inside and those outside $\mathcal{P}$ the polygon, which are used for scene boundary estimations. The different possibilities for finding a source-microphone-polygon can be seen. When using all estimated source and microphone positons marked as ○, more reflection points • might be excluded, resulting in possible differences of the scene reconstruction. In this work, the averaged estimates are used.

3.3.3 Reflector Clustering with Wall Angle Assumption

The first room inference algorithm uses the estimated wall angles ω to cluster the reflection points together. These clusters can be separated in two sub-clusters, given that opposing walls in a shoebox-shaped room have the same angle. The initial set of points is $\mathcal{P}$ with $z_{\mathcal{P}}$ describing points from the set.

In the beginning an exclusion is performed by only allowing points to remain in $\mathcal{P}$ for which the wall angle $\omega_{\mathcal{P}}$ is within ε degrees of any other points angle. The remaining points are stored in the starting set $\mathcal{C}_0$. In the ideal case of a perfect estimation of *all* reflection points of a room with four walls, a maximum of four different angles can be observed in rooms with no parallel walls, and a minimum of two different angles for rectangular rooms. In reality though, no two estimated angles might be the same.

First the densities of the wall angles for each point (corresponding to the number of other reflection points of which the angles $\omega_{\mathcal{C}_0}$ are within d_c to that of the examined point) are computed (similar to Equation 2.3.5) as

$$\rho_{\omega}(z_{\mathcal{C}_0}) = \sum_{\mathcal{C}_0} \exp\left(-\frac{d\left(\omega_{\mathcal{C}_0}, \omega'_{\mathcal{C}_0}\right)^2}{d_c^2}\right), \tag{3.3.10}$$

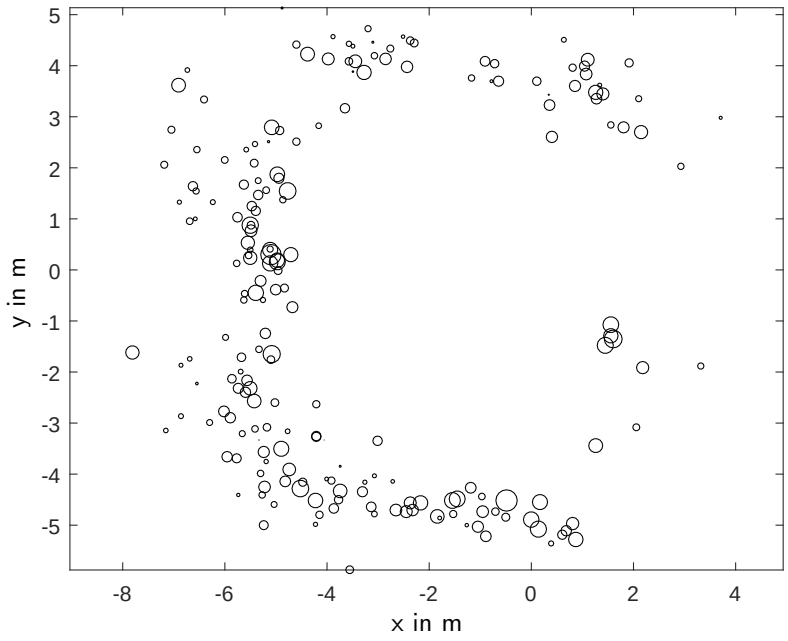


Figure 23: Example of reflection points used for room inference with the circle size indicating the underlying weight. The shape of the room can already be suspected, although no rotation is performed to align the points to the real measurement situation.

where $d\big(\omega_{\mathcal{C}_0}, \omega'_{\mathcal{C}_0}\big)$ computes distance between the wall angle $\omega_{\mathcal{C}_0}$ of the current examined point $z_{\mathcal{C}_0}$ and another point $z'_{\mathcal{C}_0}$ with angle $\omega'_{\mathcal{C}_0}$ from the set $\mathcal{C}_0$ according to Equation 2.3.3. From $\rho_\omega(z_{\mathcal{C}_0})$ the point with the highest angle density $\rho_\omega(z_{\mathcal{C}_0})$ is chosen as starting point for the first cluster $\mathcal{C}_l$ with $l = 1$ (where l denotes the cluster number) as

$$\tilde{z}_{\mathcal{C}_1} = \underset{z_{\mathcal{C}_0}}{\operatorname{argmax}} \left\{ \rho_\omega(z_{\mathcal{C}_0}) \right\}. \tag{3.3.11}$$

This first cluster $\mathcal{C}_1$ is then populated with all points from $\mathcal{C}_0$ for which the wall angle $\omega_{\mathcal{C}_0}$ is within ε degrees (usually chosen with 10°) to that of $\tilde{z}_{\mathcal{C}_1}$. For all these points in $\mathcal{C}_1$ the average wall angle $\bar{\omega}_{\mathcal{C}_1} = \frac{1}{|\mathcal{C}_1|} \sum_{\mathcal{C}_1} \omega_{\mathcal{C}_1}$ with $|\mathcal{C}_1|$ denoting the cardinality of the cluster-set $\mathcal{C}_1$ is computed. This average angle is used to rotate all points in the set such that $\bar{\omega}_{\mathcal{C}_1}$ becomes perpendicular to the real axis, which should result in the smallest possible variance of the reflection points for the respective reflector when projecting the points from the cluster onto the real axis. The rotated points are

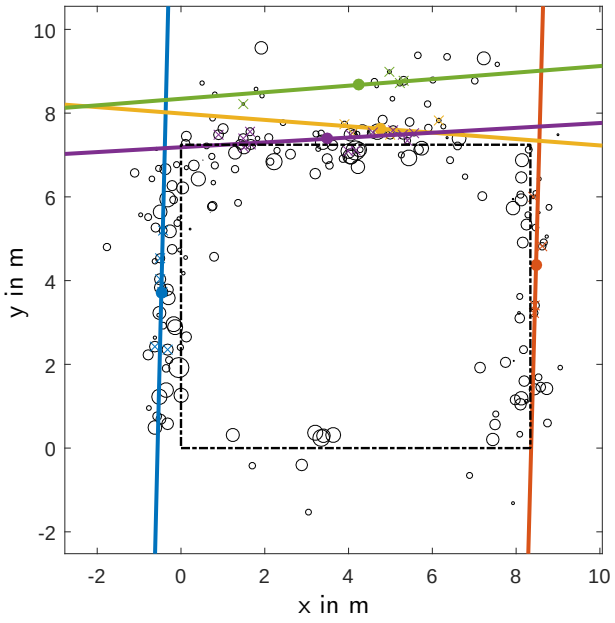


Figure 24: Room inference results when using the estimated wall angles to cluster the reflection points to find possible reflecting surfaces. The rotated reflection points $\check{z}_{\mathcal{P}}$ are shown as ○ of size corresponding to the weight. The black dashed line shows the simplified model of the room.

computed by

$$z_{\mathcal{C}_1}^{\perp} = z_{\mathcal{C}_1} \cdot e^{-i \cdot \bar{\omega}_{\mathcal{C}_1}}. \tag{3.3.12}$$

The distances between the projected points $z_{\mathcal{C}_1}^{\perp}$ are computed and all points $\mathcal{C}_1^{\perp}$ that are within a margin of $d_{\perp}$ to the point with the minimum distance to most other points (centre of the density of the projected points) are used as the final points for populating the first cluster $\hat{\mathcal{C}}_1'$ (after rotating them back to their original positions using the inverse of Equation 3.3.12).

At this point it will be checked if it is necessary to create a sub-cluster $\hat{\mathcal{C}}_1''$, containing points of a parallel wall on the opposite side of the scene. This is done by checking for another point with high density on the other side of the support spanned by the polygon points $z_{\mathcal{G}}$ (rotated and projected onto the real axis the same way as the points from the initially chosen cluster $\mathcal{C}_1$) exists. If this is the case, the sub-cluster $\hat{\mathcal{C}}_1''$ is populated similar to $\mathcal{C}_1$ (with all points within a certain margin around the point with the highest density), removing these points form the initial cluster $\hat{\mathcal{C}}_1'$.

After finding the first cluster(s), the same procedure is repeated for the points

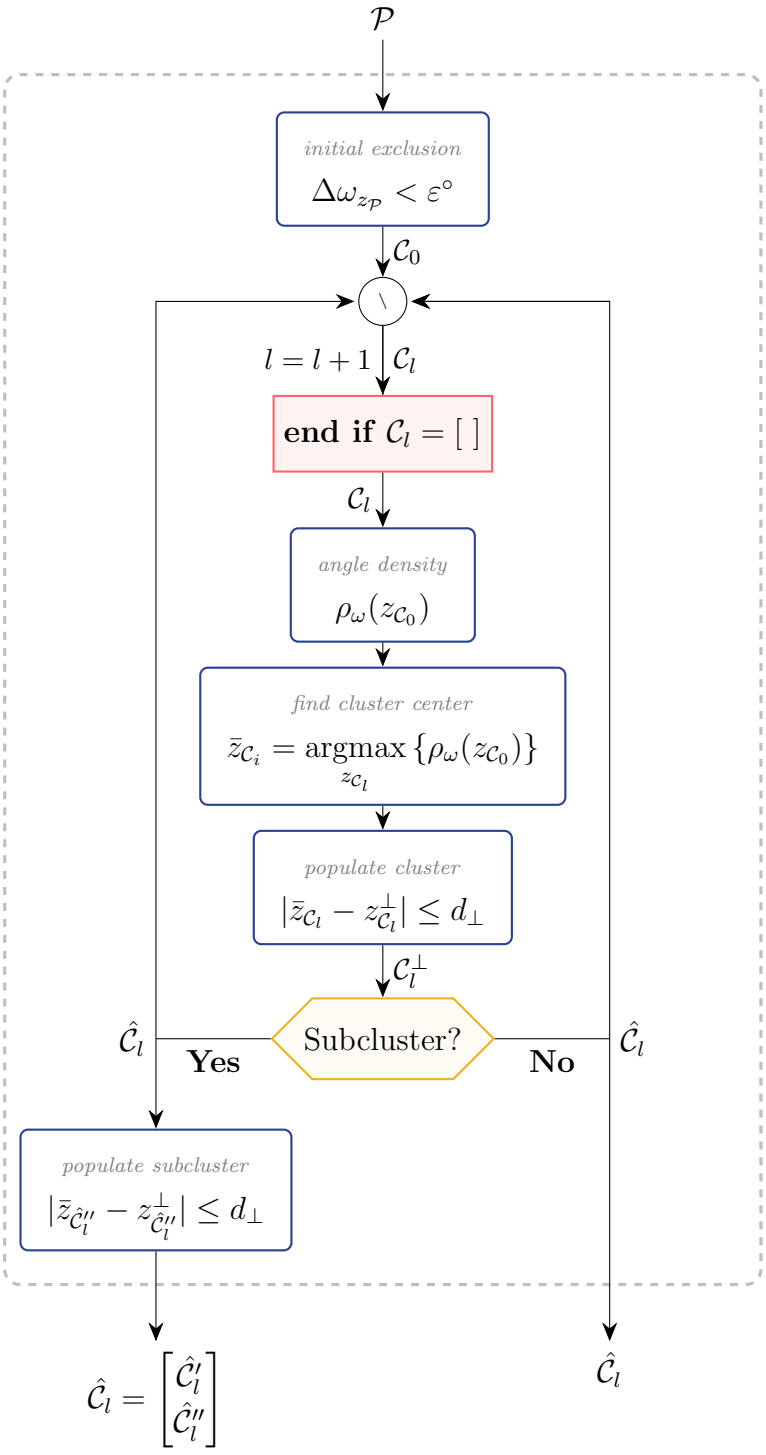


Figure 25: Flow graph for the angle clustering algorithm.

$\bar{\mathcal{C}}_0 = \mathcal{C}_0 \setminus \hat{\mathcal{C}}_1$ that were not chosen for a cluster $\hat{\mathcal{C}}_1$[8]. From these points $\bar{\mathcal{C}}_0$ again the one with the highest wall angle density is chosen to start with populating the initial second cluster $\mathcal{C}_2$ with the rest of the population performed in the same way as for the first one (rotation with $\bar{\omega}_{\mathcal{C}_2}$, population decrease to $\hat{\mathcal{C}}_2'$, possible cluster separation to $\hat{\mathcal{C}}_2''$) and then again continuing for the next cluster with all points not chosen for $\hat{\mathcal{C}}_1$ or $\hat{\mathcal{C}}_2$ until no unused points are left. It makes furthermore sense to impose a minimum population on each cluster.

Results for this algorithm are shown in Figure 24, with the different reflectors shown in different colors. A flow graph of the algorithm can be seen in Figure 25 with the set of reflection points $\mathcal{P}$ as the input and the resulting sets for the cluster $\hat{\mathcal{C}}_l'$ and a possible subcluster $\hat{\mathcal{C}}_l''$ as the output after each iteration.

3.3.4 Hough Transform for Line Detection

An elegant way of estimating reflectors from reflection points is shown by Filos [Fil13] by the Hough transform. The classical Hough transform as reviewed by Illingworth in [IK88] can be seen as a coordinate transform, initially used to detect lines in images, showing similarities to the problem of fitting a possible reflector to a set of reflection points (which are ideally located on a line). In the original algorithm for image processing the detection is performed by finding the Hough parameters $\Theta = (r, \vartheta)$ of all interesting points in the image (the parameters are depicted in Figure 2). To perform the transformation the whole new parameter space is discretized and points in the image that lie on the same parameterized line produce a maximum at the that very parameter set. The counts each parameter set receives are stored as $A(r, \vartheta)$.

In contrast to the original definition, for this algorithm the values of $A(r, \vartheta | d_{HT})$ are computed by counting the number of points within a certain allowed margin d_{HT} around the line with parameters $\Theta_{HT} = (r_{HT}, \vartheta_{HT})$. The exclusion of points that are too far away from the line can be seen as smearing of the Hough space image, depending on the cut-off threshold d_{HT}. The Hough transform used here follows the formula

$$J_{HT}(x, y | \Theta_{HT}) = x \cdot \cos \vartheta_{HT} + y \cdot \sin \vartheta_{HT} - r_{HT}, \tag{3.3.13}$$

[8] Obviously both cases $\mathcal{C}_l = \hat{\mathcal{C}}_l' \cup \hat{\mathcal{C}}_l''$ and $\mathcal{C}_l \neq \hat{\mathcal{C}}_l' \cup \hat{\mathcal{C}}_l''$ are possible, i.e. not all points initially chosen might end up in a final cluster.

with the $A(\Theta_{HT}|d_{HT})$ filled according to

$$A(\Theta_{HT}|d_{HT}) = \sum_{\mathcal{P}} \begin{cases} 1, & |J_{HT}(x,y|\Theta_{HT})| < d_{HT} \\ 0, & else \end{cases}, \tag{3.3.14}$$

which counts only reflection points in $\mathcal{P}$ that are within d_{HT} from the line described by the parameters Θ_{HT}.

To avoid the need for two-dimensional peak picking the Hough transform is performed iteratively on a reduced set of points at each iteration. After computing the first Hough transform the global maximum is found according to

$$\hat{\Theta}_{HT}^{(1)} = \underset{\Theta_{HT}}{\operatorname{argmax}} \left\{ A_{HT}(\Theta_{HT}|d_{HT}) \right\}, \tag{3.3.15}$$

which gives the parameters for the first possible reflector. The reduced set of points for the next iteration is obtained by removing all points that created that first parameter maximum. Parameters to choose are the number of iterations to perform N_w, chosen as the number of reflectors that need to be estimated.

In Figure 27 the Hough space representation for a set of randomly created points representing simulated reflection points is depicted, used to evaluate possible results achievable by the Hough transform. Four local maxima can be observed in Figure 27c (actually six can be seen due to the examined angles ranging from $-\pi$ to π, which also causes more maxima to disappear from iteration to iteration, it is sufficient to examine the parameter range $r_{HT} \geq 0$ and $0 \leq \vartheta_{HT} \leq \pi$).

The Hough space set of parameters $A(\Theta_{HT}|d_{HT})$ for reflection points estimated from data measured according to Appendix B.2 is shown in Figure 28 (maxima at each iteration indicated), showing results similar to those of the simulated points. What becomes obvious is the fact that when less reflection points are found for a certain wall that wall might not be detected, as was the case here. The wall that could not be detected is the back wall of the lecture hall that received acoustic treatment, which is the wall that would create a rectangle in Figure 28b located vertically at $x \approx 2\ m$.

What makes the use of the Hough transform attractive is the possibility of a simple inclusion of conditions concerning the angles between reflectors.

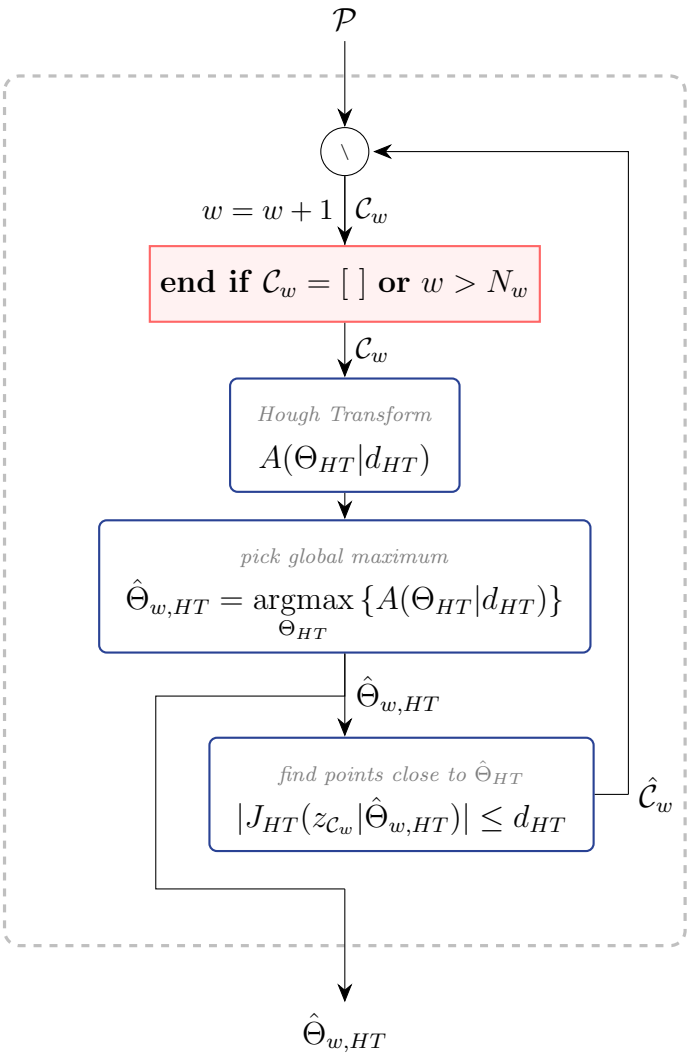
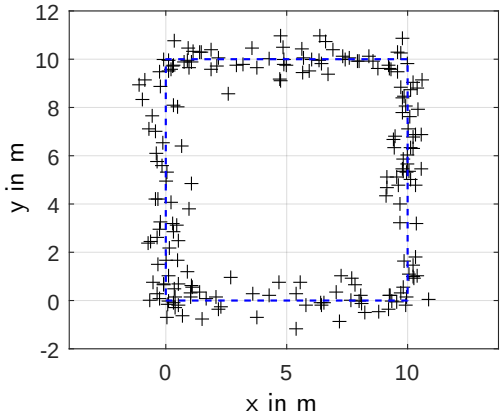
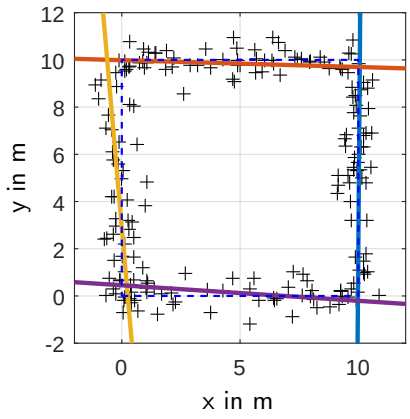


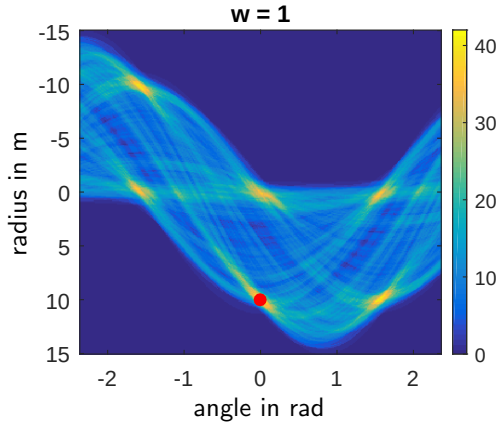
Figure 26: Flow graph for the Hough transform based algorithm.



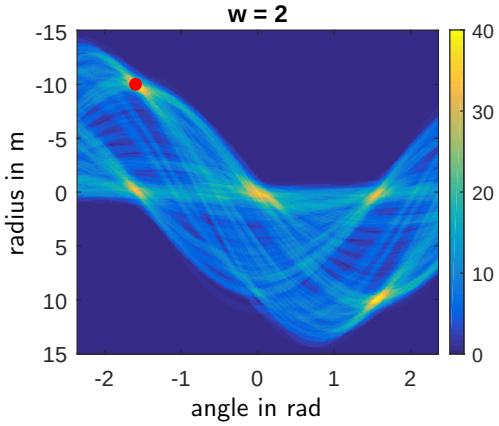
(a) simulated reflection points



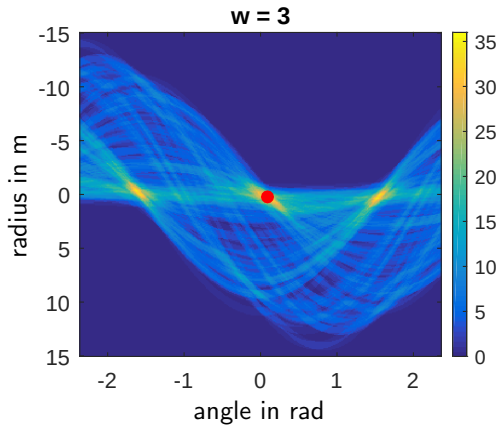
(b) estimated reflectors



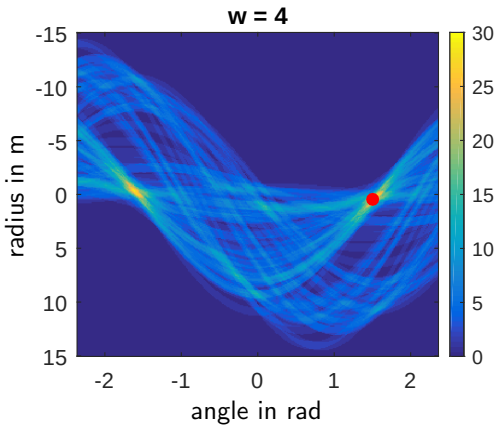
(c) iteration 1 as ▮



(d) iteration 2 as ▮

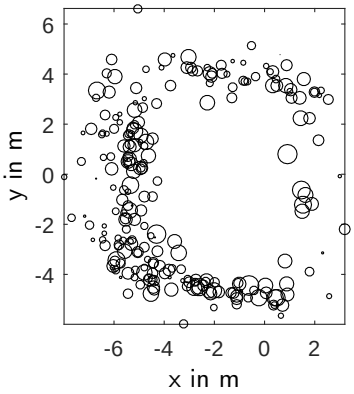


(e) iteration 3 as ▮

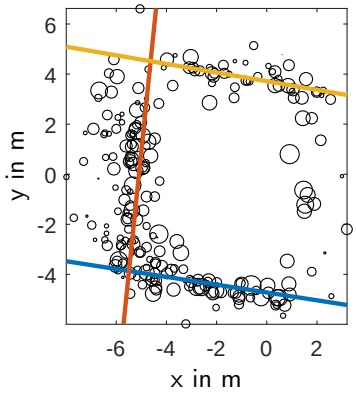


(f) iteration 4 as ▮

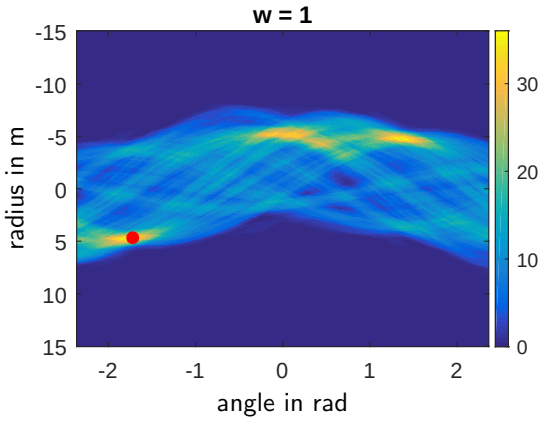
Figure 27: Hough transform algorithm applied on simulated reflection points (rectangular room with width and height of 10 m), with the corresponding representation in the Hough space showing distinct peaks at parameter spots. The periodicity is caused by the examined angle interval, the maximum of $A_{HT}^{(w)}$ for the current iteration w is marked as ●.



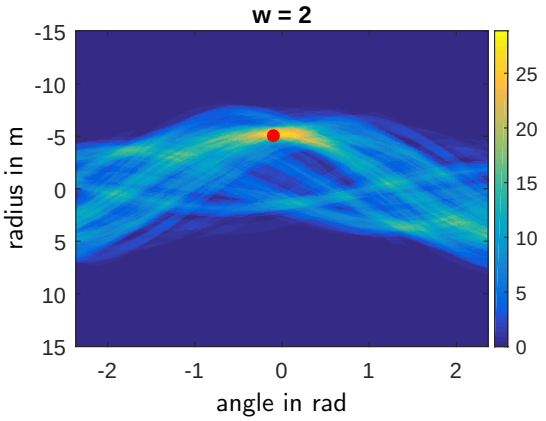
(a) estimated reflection points



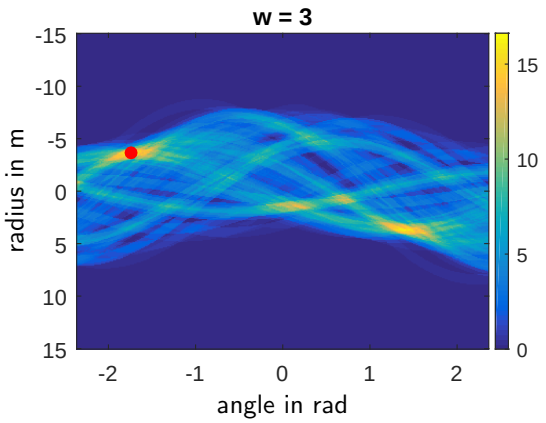
(b) estimated reflectors



(c) iteration 1 as ▮



(d) iteration 2 as ▮



(e) iteration 3 as ▮

Figure 28: Hough transform algorithm on estimated reflection points, the maximum at iterations w of $A_{HT}^{(w)}$ is marked as ●. The walls found at iteration w can be seen in Figure 28b with the wall colours indicating the iteration. As can be seen in the plots, the wall estimates achieve less counts during later iterations, which might pose a problem when having few points at disposition.

3.3.5 Principal Component Projected Histograms

Under the presumption that a room has rectangular shape another possible solution is to find the two *principle components* (PCs) of the cloud of all reflection points, the underlying assumption being that the PCs are ideally parallel to the walls of a rectangular room. By projecting the found reflection points from the set $\mathcal{P}$ onto its principal components, locations with a many projected points can be presumed to correspond to walls.

A simple way to compute the projection is to find the rotation angle of the data needed such that the principal components align to the x- and y-axis or the *real*- and *imaginary*-axis. After that the projection can be computed by simply taking the real or imaginary part when using complex notation.

From a PC $\mathbf{g}^{(q)} = \begin{pmatrix} g^{(q)}(1) & g^{(q)}(2) \end{pmatrix}^T$, where $q \in \{1, 2\}$ is the index of the principal component, the angle needed to rotate $\mathbf{g}^{(q)}$ such that the PC align to a coordinate axis is found as

$$\phi_{PC}^{(1)} = \operatorname{atan}\left(\frac{g^{(1)}(2)}{g^{(1)}(1)}\right), \tag{3.3.16}$$

which is then applied to all points from the set $\mathcal{P}$ according to

$$\breve{z}_{\mathcal{P}} = z_{\mathcal{P}} \cdot e^{-i \cdot \phi_{PC}^{(1)}}, \tag{3.3.17}$$

with the same rotation used on the source and microphone estimates (stored in the sets $\mathcal{R}$ and $\mathcal{S}$) resulting in $\breve{z}_{\mathcal{R}}$ and $\breve{z}_{\mathcal{S}}$ respectively. Since the principal components are always orthogonal to each other only one rotation angle is needed. From these points two weighted histograms

$$h_{Re}(b) = \sum_{\mathcal{P}} w_{PC}(\breve{z}_{\mathcal{P}}) \cdot \Pi_{b,\Delta_B}\left(\operatorname{Re}\left\{\breve{z}_{\mathcal{P}}\right\}\right) \tag{3.3.18}$$

$$h_{Im}(b) = \sum_{\mathcal{P}} w_{PC}(\breve{z}_{\mathcal{P}}) \cdot \Pi_{b,\Delta_B}\left(\operatorname{Im}\left\{\breve{z}_{\mathcal{P}}\right\}\right) \tag{3.3.19}$$

can be computed where $w_{PC}(\breve{z}_{\mathcal{P}})$ is the weight of the reflection point $\breve{z}_{\mathcal{P}}$ and $\Pi_{n,N}(\cdot)$ is the selection function defined earlier as

$$\Pi_{n,N}(x) = \begin{cases} 1, & N \cdot \left(n - \frac{1}{2}\right) < x \leq N \cdot \left(n + \frac{1}{2}\right) \\ 0, & else \end{cases} \tag{3.3.20}$$

The sums are computed over all points in the set $\mathcal{P}$ with $b \in \{0, 1, \ldots, N_b - 1\}$ representing the bin index and Δ_B the bin spacing. N_b is the number of histogram bins.

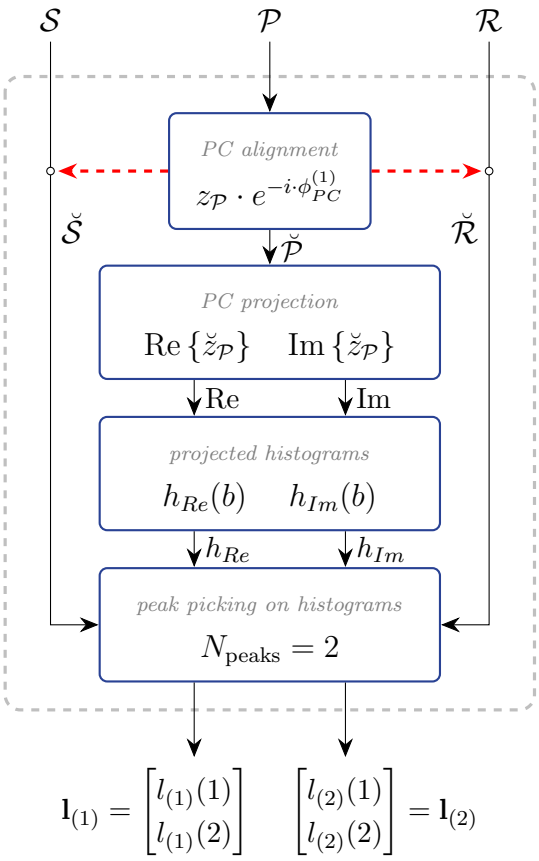


Figure 29: Flow graph for the PC projection algorithm.

The positions of two maxima $\mathbf{l}^{(1)} = \begin{pmatrix} l^{(1)}(1) & l^{(1)}(2) \end{pmatrix}^T$ on the first and $\mathbf{l}^{(2)} = \begin{pmatrix} l^{(2)}(1) & l^{(2)}(2) \end{pmatrix}^T$ on the second principal component that are found outside of the support of the projected source and microphone points $z_{\mathcal{R}}$ and $z_{\mathcal{S}}$ are taken as estimated positions of reflectors. The orientation of these reflectors is then perpendicular to the respective PC (or parallel to the other PC).

The support of the PC projected microphone points is the interval $\mathbb{S}_{\mathrm{Re}\{\mathcal{R}\}} = \left[\min \mathrm{Re}\{\breve{z}_{\mathcal{R}}\}, \max \mathrm{Re}\{\breve{z}_{\mathcal{R}}\}\right]$ and $\mathbb{S}_{\mathrm{Im}\{\mathcal{R}\}} = \left[\min \mathrm{Im}\{\breve{z}_{\mathcal{R}}\}, \max \mathrm{Im}\{\breve{z}_{\mathcal{R}}\}\right]$ (i.e. the outermost of the microphone points projected on the respective PC) with equivalent definitions for the source points $z_{\mathcal{S}}$. The joint support of the projected microphone and source points are then $\mathbb{S}_{\mathrm{Re}} = \mathbb{S}_{\mathrm{Re}\{\mathcal{R}\}} \cup \mathbb{S}_{\mathrm{Re}\{\mathcal{S}\}}$ and $\mathbb{S}_{\mathrm{Im}} = \mathbb{S}_{\mathrm{Im}\{\mathcal{R}\}} \cup \mathbb{S}_{\mathrm{Im}\{\mathcal{S}\}}$ respectively.

A flow graph of proposed the algorithm can be seen in Figure 29. Exemplary results for data from the second measurement (described in Appendix B.2) can be seen in Figure 30, showing the weighted histograms of the reflection points $\mathcal{P}$ projected onto

the PCs outside of the source and microphone supports (denoted by red $\times$) displayed in Figures 30b and 30c. The peaks are then picked outside of the support $\mathbb{S}_{\mathrm{Re}}$ and $\mathbb{S}_{\mathrm{Im}}$ are indicated by $\blacklozenge$ with colors matching the reflector. In Figure 30a the colors of the reflector lines are matched to the colors of the peaks in the histograms of the projected reflection points.

3.3.6 Rectangular Fit

The last method assumes that, similar to linear regression where a set of points is modelled by the line of best fit, the reflection points can be fitted with a rectangle which in turn can be approximated using the slightly modified equation of an ellipse

$$\left(\frac{x-x_0}{a}\right)^{2\eta} + \left(\frac{y-y_0}{b}\right)^{2\eta} = 1. \tag{3.3.21}$$

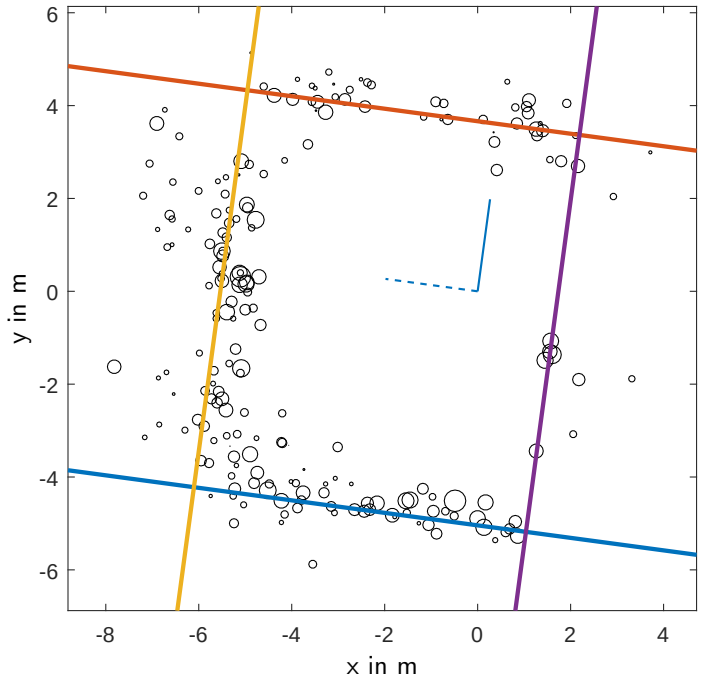
Using $\eta = 1$ results in an ellipse with width along the x- and y-axis equal to $2 \cdot a$ and $2 \cdot b$ respectively, centred around (x_0, y_0). By increasing η the ellipse can be transformed into a rectangle with slightly rounded corners and width along x and y-axis remaining equal to $2 \cdot a$ and $2 \cdot b$ with the same centre point as the original ellipse. Examples for different exponents η are shown in Figure 31. For values $\eta \geq 2$ Equation 3.3.21 will imitate a rectangle close enough for the task at hand. The slightly rounded edges that remain do not pose a problem since the estimated reflection points ususlly do not follow the actual room geometry that accurately either (see in Figure 33).

A measure for the quality of the fit for each reflection point can be computed with the help of Equation 3.3.21 as

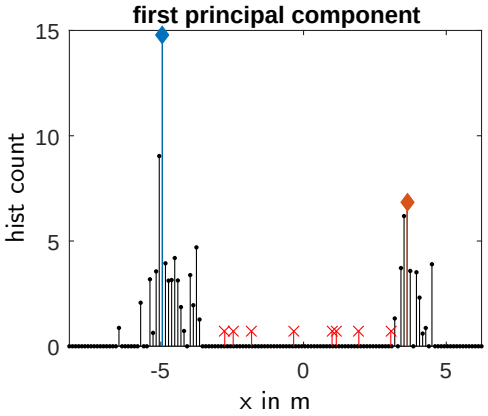
$$D_{\square}(x, y|\boldsymbol{\Theta}_{\square}) = \left(\frac{x-x_0}{a}\right)^{2\eta} + \left(\frac{y-y_0}{b}\right)^{2\eta} - 1, \tag{3.3.22}$$

which will ideally be minimal when all points are located on the rectangle with parameters $\boldsymbol{\Theta}_{\square} = (\boldsymbol{\Theta}_0, \boldsymbol{\Theta}_+) = (x_0, y_0, a, b)$ with $\boldsymbol{\Theta}_0$ containing the centre and $\boldsymbol{\Theta}_+$ the width parameters. The problem with Equation 3.3.22 is that points that are located inside the rectangle with parameters $\boldsymbol{\Theta}_{\square}$ are weighted less (i.e. negative) than to those outside the rectangle which would result in a rectangle that *encloses* all reflection points when performing the fitting, which is not the desired result. This problem can be overcome by the use of the cost function

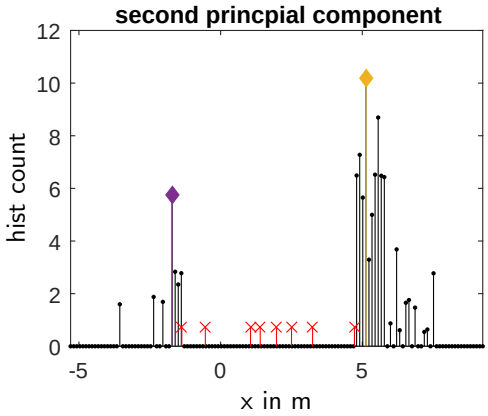
$$J_{\square}(x, y|\Theta_{\square}) = \sum_{\mathcal{P}} \operatorname{atan}\left(D_{\square}(x, y|\Theta_{\square}) - \frac{\pi}{8}\right)^2, \tag{3.3.23}$$



(a) estimated walls (▮,▮,▮ and ▮) and the first (**solid**) and second (**dashed**) principal components indicating the positive axes of the histograms

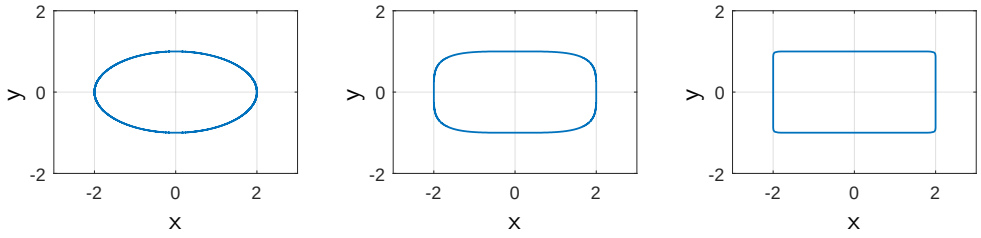


(b) histogram of $\check{z}_{\mathcal{P}}$ projected onto the first principal component



(c) histogram of $\check{z}_{\mathcal{P}}$ projected onto the second principal component

Figure 30: Room inference when projecting the estimated reflection points $\check{z}_{\mathcal{P}}$ onto the principal components of those points. The support of polygon formed by the microphone and sources (×) projected onto the principal components (using Equation 3.3.17) are excluded from the histograms, resulting in two distinct peaks that can be found easily, marked in the same colour as the corresponding reflector.



(a) $\eta = 1$. **(b)** $\eta = 2$. **(c)** $\eta = 15$.

Figure 31: Transforming an ellipse to a rectangle by changing η, with $a = 2$ and $b = 1$ and $x_0 = y_0 = 0$.

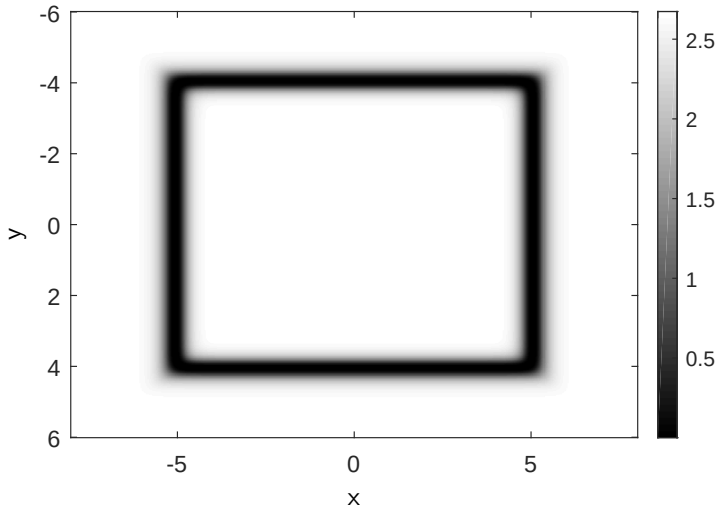


Figure 32: Cost function $J_{\square}(x, y|\boldsymbol{\Theta}_{\square})$ for different points (x, y) and a fixed set of parameters $\boldsymbol{\Theta}_{+} = (5, 4)$ and $\boldsymbol{\Theta}_{0} = (0, 0)$.

which implements an equal weighting inside and outside of the rectangle corresponding to parameters $\boldsymbol{\Theta}_{\square}$, with the sum computed over all points in $\mathcal{P}$. The error space of the cost function is displayed in Figure 32. It is trivial that the optimal parameters will result in the maximum possible points being located close to the edges of the rectangle (the dark area in Figuer 32), as well as an increasing quality of the fit with increasing accuracy of the positions of the estimated reflection points.

The optimization was performed in Matlab using `fminsearch` leading to good results within negligible time. The final result on data from the measurement conducted in the lecture hall (Appendix B.2) are presented in Figure 33.

Initialization

Since a good initialization is always an advantage in terms of convergence, the centre point $\boldsymbol{\Theta}_0$ is initialized with the point inside the source-microphone polygon that is farthest away from all points in $\mathcal{P}$. The distance to all points can be calculated using

$$d\left(\boldsymbol{\Theta}_0\right) = \sum_{\mathcal{P}} ||\boldsymbol{\Theta}_0 - \breve{\mathbf{z}}_{\mathcal{P}}||_2, \tag{3.3.24}$$

where $\boldsymbol{\Theta}_0 = \begin{pmatrix} x_0 & y_0 \end{pmatrix}^T$ is a vector describing the ellipse centre and $\breve{\mathbf{z}}_{\mathcal{P}}$ a reflection point from the set $\mathcal{P}$ in vector form (the sum is computed over all points in the set). As described in Section 3.3.5 by Equation 3.3.16, the points in $\mathcal{P}$ are rotated to align the principal components to the coordinate axes to eliminate the need to find a rotation parameter for the rectangle (if the reflection points are accurate enough). This rotation could also be seen as an initialization of the rotation angle. The possible introduction of a specific rotation-parameter $\vartheta_{\circlearrowleft}$ for optimization alongside the parameters $\boldsymbol{\Theta}_0$ and $\boldsymbol{\Theta}_+$ is described right after the initialization.

The centre initialization inside of the source-microphone polygon can then be found using gradient descent or a search over the complete polygon space on the formula

$$\boldsymbol{\Theta}_0^{(init)} = \underset{\boldsymbol{\Theta}_0}{\operatorname{argmin}} \left\{-d\left(\boldsymbol{\Theta}_0\right)\right\}. \tag{3.3.25}$$

To find the starting value for $\boldsymbol{\Theta}_+$, the following equations give reasonable values:

$$a^{(init)} = \frac{\max(\operatorname{Re}\left\{\breve{z}_{\mathcal{G}}\right\}) - \min(\operatorname{Re}\left\{\breve{z}_{\mathcal{G}}\right\})}{2} \tag{3.3.26}$$

$$b^{(init)} = \frac{\max(\operatorname{Im}\left\{\breve{z}_{\mathcal{G}}\right\}) - \min(\operatorname{Im}\left\{\breve{z}_{\mathcal{G}}\right\})}{2} \tag{3.3.27}$$

Therein $\breve{z}_{\mathcal{G}}$ are the points that create the source-microphone polygon aligned to the PCs as described in Section 3.3.5 by Equation 3.3.16, resulting in the width initialization of

$$\boldsymbol{\Theta}_+^{(init)} = \begin{pmatrix} a^{(init)} & b^{(init)} \end{pmatrix}^T. \tag{3.3.28}$$

This ensures that the optimization process starts from the inside of points in $\mathcal{P}$ and expands the rectangle outwards, which might prevent accidental stopping when finding a local minimum caused by an accumulation of inaccurate reflection points.

Additional Rotation Parameter

For the case that the alignment using the principal components of the estimated

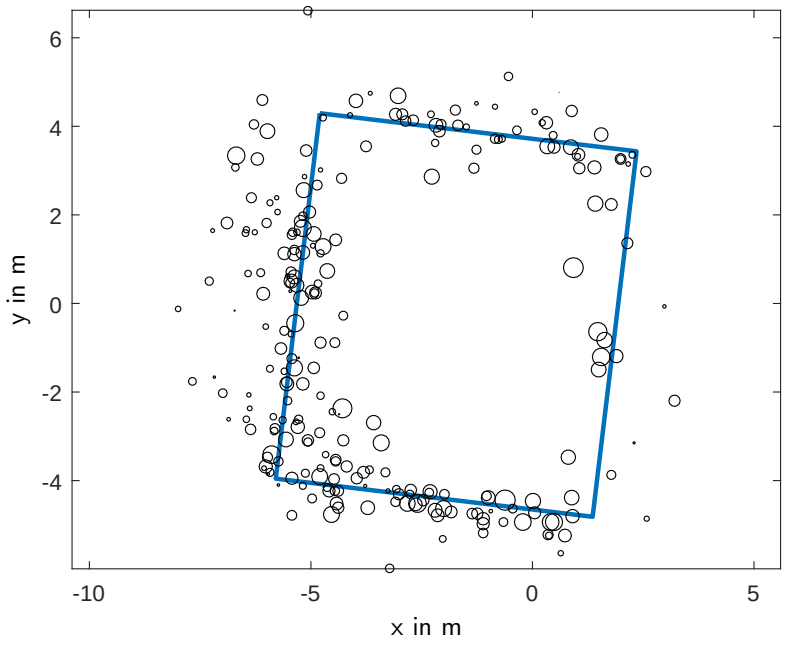


Figure 33: Result of room inference when using the rectangular fit without the additional rotation parameter, with the estimated rectangle indicated in **blue** and the estimated reflection positions included.

reflection points is not sufficient an additional rotation parameter $\vartheta_{\circlearrowleft}$ can be introduced to the rectangular fit. The easiest way to do this is to perform the normal rectangular fit algorithm and then find the optimal rotation of the reflection points to fit the estimated rectangle, continuing in an iterative fashion and initializing the next iteration with the optimal parameters $\vartheta_{\circlearrowleft}$ and $\boldsymbol{\Theta}_{\square}$ from the prior one. As stopping criterion either a fixed number of iterations $p_{\max}$ or a bound on the magnitude of the rotation angle $\vartheta_{\circlearrowleft}^{(p)}$ can be used[9]. Due to allowing only small rotations at each iteration, the overall rotation $\tilde{\vartheta}_{\circlearrowleft}$ (in addition to the PC alignment) applied can be found by summing over the rotation angles of each iteration according to

$$\tilde{\vartheta}_{\circlearrowleft} = \sum_{p} \vartheta_{\circlearrowleft}^{(p)}. \tag{3.3.29}$$

The rotation allowed at each iteration can be bounded within a small range around zero, for example $\vartheta_{\circlearrowleft} \in \left[-\frac{\pi}{10}, \frac{\pi}{10}\right]$. The best rectangular fit in terms of the centre points $\boldsymbol{\Theta}_0$ and rectangle height and width $\boldsymbol{\Theta}_+$ is found again using the cost function

[9]The rotation angle $\vartheta_{\circlearrowleft}^{(p)}$ usually approached zero with advancing iterations as $\lim_{p \to p_{\max}} \vartheta_{\circlearrowleft}^{(p)} \approx 0$

$J_{\square}(x, y|\Theta_{\square}, \vartheta_{\circlearrowleft})$ followed by the optimization for the rotation using

$$J_{\circlearrowleft}(z|\boldsymbol{\Theta}_{\square}, \vartheta_{\circlearrowleft}) = \sum_{\mathcal{P}} \operatorname{atan}\left(D_{\circlearrowleft}(z|\boldsymbol{\Theta}_{\square}, \vartheta_{\circlearrowleft}) - \frac{\pi}{8}\right)^2, \tag{3.3.30}$$

with

$$D_{\circlearrowleft}(z|\boldsymbol{\Theta}_{\square}, \vartheta_{\circlearrowleft}) = \left(\frac{\operatorname{Re}\left\{z \cdot e^{i\vartheta_{\circlearrowleft}}\right\} - x_0}{a}\right)^{2\eta} + \left(\frac{\operatorname{Im}\left\{z \cdot e^{i\vartheta_{\circlearrowleft}}\right\} - y_0}{b}\right)^{2\eta} - 1, \tag{3.3.31}$$

and $z = x + iy$ being the reflection points in complex coordinates. Finding the optimal rotation angle $\vartheta_{\circlearrowleft}$ can again be performed by picking the parameter corresponding to the minimum of the cost function $J_{\circlearrowleft}(z|\boldsymbol{\Theta}_{\square}, \vartheta_{\circlearrowleft})$ from the chosen bounded parameter range.

3.3.7 Brief Reflector Localization Results

In this section only a short presentation of a few figures will be given with the implementations still fresh in mind. A more thorough examination of the algorithm performance can be found in Chapter 4. The symbols used in the figures are shown in Table 3.

Table 3: Symbols used in the estimated maps.

$\hat{\mathbf{s}}_{ij}$	□	calibration source positions j estimated from microphone i
$\bar{\mathbf{s}}_j$	●	calibration source position j averaged over all microhpone estimates
$\mathbf{s}_j$	□	measured position of calibration source j
$\hat{\mathbf{r}}_i$	■	estimated position of microphone i
$\mathbf{r}_i$	×	measured position of microphone i

A problem when comparing the results of found reflectors are the different possibilities of alignment of the estimated to the real (measured) results and inherent measurement errors of microphone, source and reflector positions. Since the aim was first to find the source and microphone positions (i.e. perform an initialization of the distributed microphone array), the microphone positions will be aligned in such a way that the estimated reference source is estimated perfectly ($\hat{\mathbf{s}}_A = \mathbf{s}_A$) and such that the look direction from the reference source $\mathbf{s}_A$ to the reference microphone $\mathbf{r}_1$ corresponds to the measured one. This would also result in the fewest measurements for performing the alignment by hand when using the system later on. Alignment was always

performed on the whole estimated scene, i.e. the coordinate system of the estimated scene was aligned with the coordinate system of the measured real scene.

When comparing the results the different algorithms show their intended field of application (rectangular rooms or arbitrary rooms). For rectangular rooms, *PC projection* and *rectangular fit* presented in Figures 34 and 35 are the most promising, since they force right angles and allow to finding a correct shape even when only few reflection points are detected for a certain wall. Additionally, the rectangular fit seems to perform better because of an inherent averaging of the available reflection points by using a continuous cost function, compared to the PC projection method which uses histograms (i.e. discretization) and is thus dependent on the chosen spacing of the histogram bins. The rectangular fit on the other hand needs good initializations of the centre of the rectangle and width along the x and y axes to ensure convergence, although this can be achieved rather elegantly as described in Section 3.3.6.

The other two reconstruction techniques, *Hough transform* and *angle clustering* with results shown in Figures 36 and 37, do not impose any restrictions. For the case that the reflection points and therefore the computed corresponding reflector angle can be assumed to be correct, angle clustering will also result in a good estimate of the real scene boundary. When the reflection point moves farther away from the real reflection point the wall angles obviously diverge from the correct value as well and the angle clustering method might no longer yield correct clusters. In this case or when many reflection points of unknown accuracy are available, the Hough transform might lead to better results with less effort. Additionally it allows for an easy inclusion of restrictions on the parameters with respect to those of the other walls.

For the first measurement, PC projection and angle clustering implementations are useful, since only a single reflector is present in this scene. The Hough transform does not perform that well due to the small number of reflection points that could be estimated in this case.

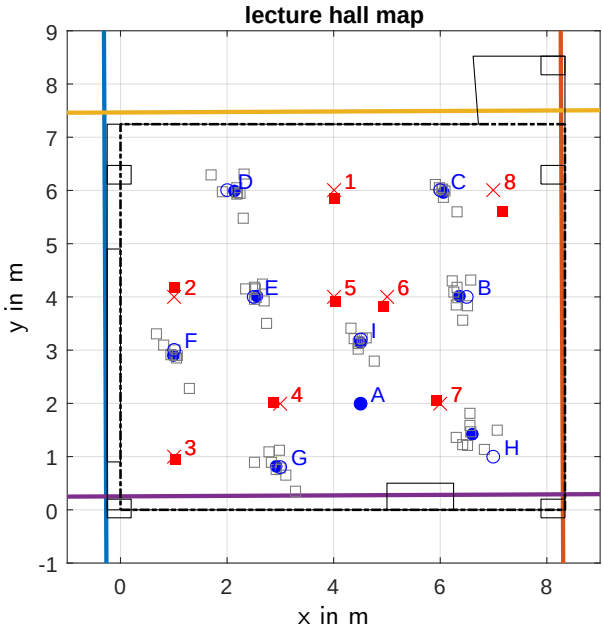


Figure 34: PC projected histograms performed on data of the second measurement (see Appendix B.2).

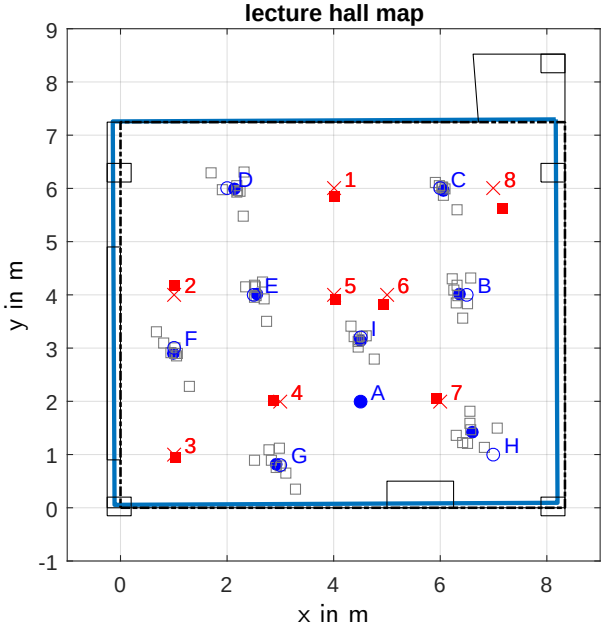


Figure 35: Rectangular fit performed on data of the second measurement (see Appendix B.2) without the additional rotation.

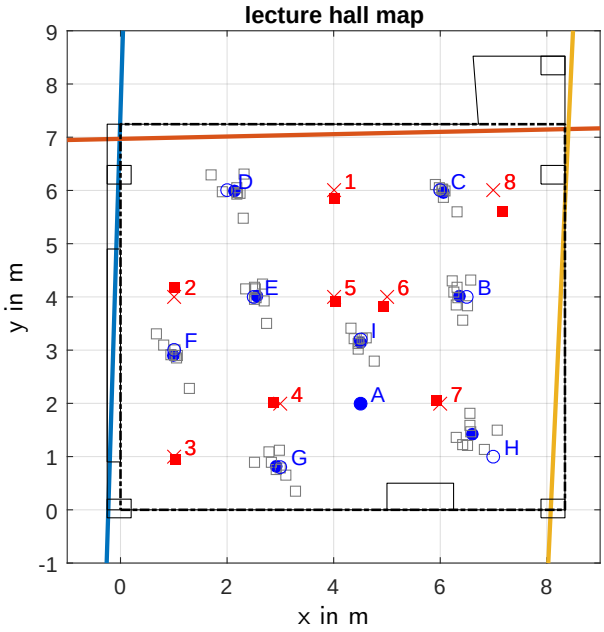


Figure 36: Hough transform for line detection performed on data of the second measurement (see Appendix B.2).

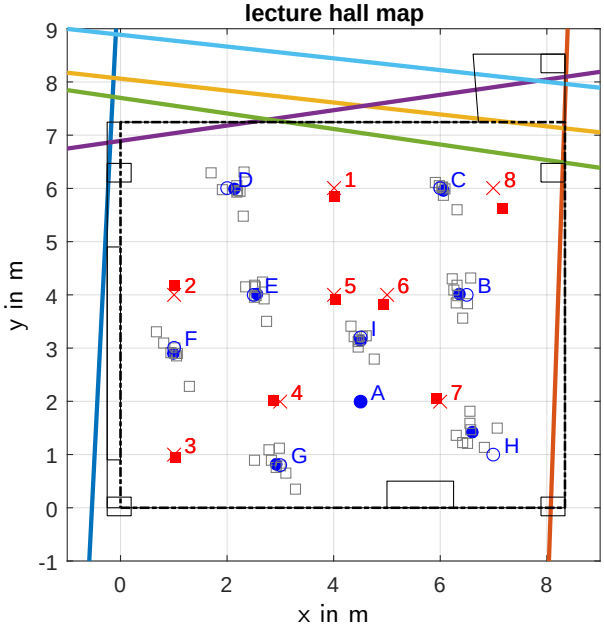
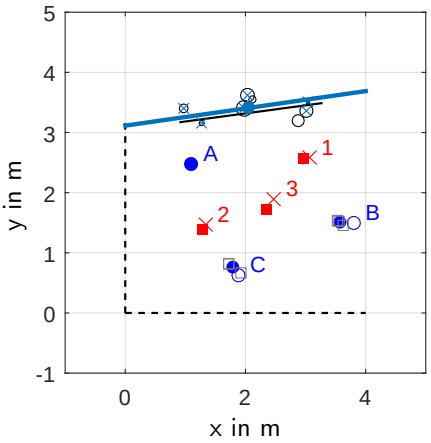
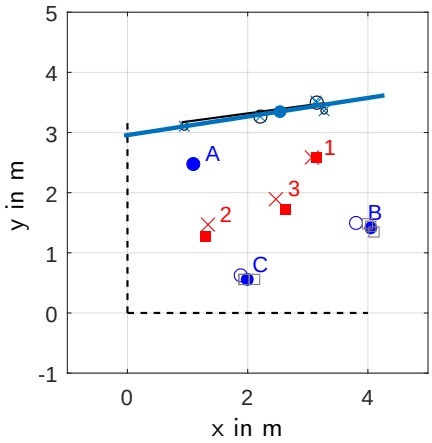


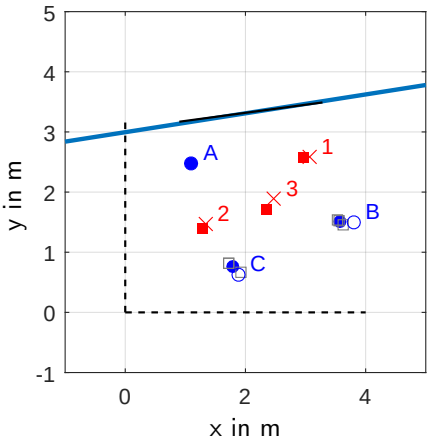
Figure 37: Wall angle clustering performed on data of the second measurement (see Appendix B.2).



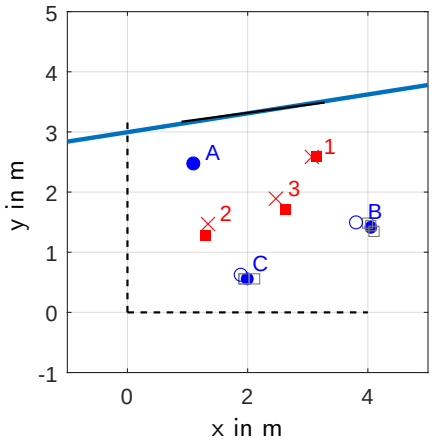
(a) wall angle clustering, rep.1



(b) wall angle clustering, rep.2



(c) PC projection, rep.1



(d) PC projection, rep.2

Figure 38: Results of the self-calibration problem using the data form the first measurement shown for two different repetitions (see Appendix B.1).

3.4 Floor and Ceiling Estimation

To estimate the distances from the microphones to the floor and the ceiling only timing information is used because of the inaccurate performance of the DOA estimation in terms of elevation angles. These estimates are obtained assuming that the locations of sources and microphones are already available. Simplification is possible by knowledge that the microphones and sources are located on a plane of constant height. Computations are shown for a single source microphone pair (i, j).

When only the TDOAs are known possible reflection points are constrained on an ellipse with a microphone i and source j in the focal points. From the point of view of the microphone, the ellipse z_{ell} (located in the xz-plane as shown in Figure 39) can be described by a rotating complex phasor

$$z_{\text{ell}}(\alpha, \ell) = r_e(\Delta \underline{t}_{i,j}^{(\ell)}, \alpha) \cdot e^{i \cdot \alpha}. \tag{3.4.1}$$

with radius given by the function

$$r_e(\Delta \underline{t}_{i,j}^{(\ell)}, \alpha) = \frac{2b \cdot \Delta \underline{t}_{i,j}^{(\ell)} + (\Delta \underline{t}_{i,j}^{(\ell)})^2}{2(b + \Delta \underline{t}_{i,j}^{(\ell)}) - 2b \cos \alpha}, \tag{3.4.2}$$

dependent on the estimated reflection TDOAs $\Delta \underline{t}_{i,j}^{(\ell)}$ and the (in this case unknown) elevation DDOAs α.

From this complex ellipse the distance from the source-microphone plane to ceiling and floor can be computed easily as either the minimum or maximum of the imaginary part

$$\hat{h}_{c,f}(\ell) = \max_{\alpha} \{\text{Im}\{z_{\text{ell}}(\alpha, \ell)\}\} \tag{3.4.3}$$

with only one explicitly computed due to the symmetry of the ellipse. The found distance will be assigned to either the ceiling or the floor, depending on a score derived from the elevation DOA estimates. This score s_θ is computed as a weighted sum over all elevation DOAs of the frame histograms $\mathbf{q}_{i,j}(m)$ around the found reflection TOA (for example using ± 2 samples as in Equation 3.4.4). If the score s_θ is larger than zero the reflection is assigned to the ceiling, if it is below zero it is assigned to the floor (a score $s_\theta = 0$ is ignored). The score is computed by

$$s_\theta(\ell) = \text{sgn}\left(\sum_{m=t(\ell)-2}^{t(\ell)+2} \mathbf{q}_{i,j}^T(m) \cdot \mathbf{b}_q \right) \tag{3.4.4}$$

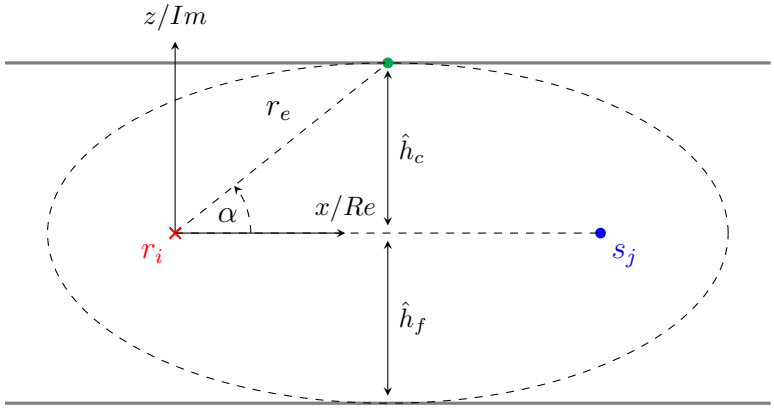


Figure 39: Ellipse for estimating floor and ceiling distances $\hat{h}_f$ and $\hat{h}_c$.

where $t(\ell)$ is the TOA of the reflection examined and $\mathbf{q}_{i,j}(m)$ the corresponding elevation DOA histogram in vector form for which the bin centres are stored in the column vector $\mathbf{b}_q$.

Performing this for all TDOAs from the set of $\mathcal{P}_{c,f}$, all found distances $\hat{h}_{c,f}(\ell)$ can be assigned to either floor or ceiling following

$$\hat{h}_{c,f}(\ell) : \begin{cases} \hat{h}_c(u_c) = \hat{h}_{c,f}(\ell), & s_\theta > 0 \\ \hat{h}_f(u_f) = \hat{h}_{c,f}(\ell), & s_\theta < 0 \end{cases} \quad \forall\ \ell, \tag{3.4.5}$$

where u_c and u_f are indices to store the heights after assigning them to floor or ceiling (they are increased after an assignment). After performing this for all candidates found for all source and microphone combinations the estimated distances to floor and ceiling are found by averaging according to

$$\bar{h}_c = \frac{1}{N_c} \sum_{u_c=1}^{N_c} \hat{h}_c(u_c) \tag{3.4.6}$$

$$\bar{h}_f = \frac{1}{N_f} \sum_{u_f=1}^{N_f} \hat{h}_f(u_f). \tag{3.4.7}$$

When floor and ceiling are not parallel to the plane of sources and microphones a common tangent approach similar to the on used by Filos [Fil13] has to be used instead of averaging over the values for the axis of the ellipse giving the height.

A diagram showing an example for an estimated ellipse with the floor and ceiling distances marked is shown in Figure 39.

4 Scene Reconstruction Results

In this chapter the results of the room inference will be presented in pictures as well as numbers. All results stem from actual measurements, performed as described in Appendices B.1 and B.2. The first examined room was designed specifically for acoustic measurements (with absorptive walls), the other is an ordinary lecture hall containing different types of surfaces (a blackboard, an acoustically treated wall, a glass front, and an ordinary wall as well as a wooden floor and a suspended ceiling).

The results for each problem are presented separately, starting with the self-calibration task and continuing with the parts that need these locations, the room inference and the height estimation. Each section will give a description of the used quality measures. The symbols to indicate microphone and calibration source positions used in all following plots are explained in Table 4.

Table 4: Symbols used in the estimated maps.

$\hat{\mathbf{s}}_{ij}$	□	calibration source positions j estimated from microphone i
$\bar{\mathbf{s}}_j$	●	calibration source position j averaged over all microhpone estimates
$\mathbf{s}_j$	□	measured position of calibration source j
$\hat{\mathbf{r}}_i$	■	estimated position of microphone i
$\mathbf{r}_i$	×	measured position of microphone i

4.1 Self-calibration

4.1.1 Localization Error Measures

To measure the accuracy of the estimated source locations, they are evaluated by computing the *mean position error* (MPE) defined as

$$\varepsilon_{\mathbf{s}} = \frac{1}{N_s} \sum_{N_s} ||\bar{\mathbf{s}}_j - \mathbf{s}_{j,opt}||_2 \tag{4.1.1}$$

where $\bar{\mathbf{s}}_j$ is the estimated source position (computed as the average of all individual estimates) of source j in Cartesian coordinates and $\mathbf{s}_{j,opt}$ the corresponding

real/measured position in Cartesian coordinates. The same measure is used for the microphone localization error, computed as

$$\varepsilon_{\mathbf{r}} = \frac{1}{N_r} \sum_{N_r} ||\hat{\mathbf{r}}_i - \mathbf{r}_{i,opt}||_2, \tag{4.1.2}$$

again yielding a single value accuracy measure for all estimated microphone points of one repetition.

Results are furthermore combined in two values, the *mean* $\mu(x)$ and the *standard deviation* $\sigma(x)$ computed as

$$\mu(x) = \frac{1}{N_e} \sum_{e=1}^{N_e} x_e \tag{4.1.3}$$

$$\sigma(x) = \sqrt{\frac{1}{N_e} \sum_{e=1}^{N_e} \big|x_e - \mu(x)\big|^2} \tag{4.1.4}$$

where the variable for which the mean and standard deviation are computed is inserted for x_e (e representing different repetitions).

4.1.2 Results when using all Microphones and Calibration Sources

Figure 40 shows the results for the self-calibration taks for data from the first measurement. It can be observed that averaging over the three separate source position estimates $\hat{\mathbf{s}}_{i,j}$ results in final estimates $\bar{\mathbf{s}}_j$ that fit the real points $\mathbf{s}_j$ closely. The overall rotation of the estimated results is computed such that the rotation angle of the first microphone is identical to the one of the real (measured) microphone direction in the coordinate system used for the room model, i.e. the line connecting reference source[10] A and microphone 1 will be parallel in the estimated and the measured models, with the estimated points shifted such that the reference sources is localized optimally, i.e. $\hat{\mathbf{s}}_A \equiv \mathbf{s}_A$.

As described in Appendix B.1, the positions of the microphones and sources were hand-measured before the acoustic measurement. The hand-measured microphone positions should therefore be accurate, while the accuracy of the actual source positions (the point where the clap occurs, i.e. the hands meet) cannot be known for

[10]The source points were denoted with capital letters $j = \{A, B, \ldots, I\} \equiv \{1, 2, \ldots, 9\}$ to make them better distinguishable from microphone points in the plots, with the corresponding numbers used in the equations and the letters in the plots.

sure. Therefore the position of the hand-measured source points might be considered more as a general area of the source position.

Self-calibration results for the first repetition of the second measurement (Appendix B.2) can be seen in Figure 41, exhibiting a similar accuracy as the results for the first measurement, although the spread of the distinct estimates of each source from the different microphones is larger than in the first measurement.

Numerical results for the MPEs for the first and second measurement are presented in Table 5 showing similar results in mean and standard deviation in both rooms.

Table 5: Localization error for the estimated microphone and source positions from measurements 1 and 2, each using the full number of microphones and sources available, indicated by N_i and N_j respectively.

	Measurement 1 $N_i = 3$, $N_j = 3$		*Measurement 2* $N_i = 8$, $N_j = 9$	
rep.	$\varepsilon_{\mathbf{r}}$ in m	$\varepsilon_{\mathbf{s}}$ in m	$\varepsilon_{\mathbf{r}}$ in m	$\varepsilon_{\mathbf{s}}$ in m
1	0.2579	0.1351	0.1666	0.1339
2	0.1595	0.2063	0.1557	0.2390
3	0.0830	0.0748	0.1276	0.3307
4	0.1174	0.0717	0.1162	0.1401
5	0.3232	0.2352	0.0966	0.1572
6	0.2945	0.3519	0.3982	0.2770
μ	**0.2059**	**0.1792**	**0.1768**	**0.2130**
σ	**0.0944**	**0.1078**	**0.1114**	**0.0819**

A superposition of the results from more repetitions is shown in Figure 42 for the first measurement, combining six clap repetitions, and in Figures 43 and 44 for the second measurement, combining five repetitions. The plots show the respective mean estimate over all separate estimates (indicated the symbols from Table 4) as well as the error ellipses which indicate the 70% confidence interval of the estimated points (sources and microphones).

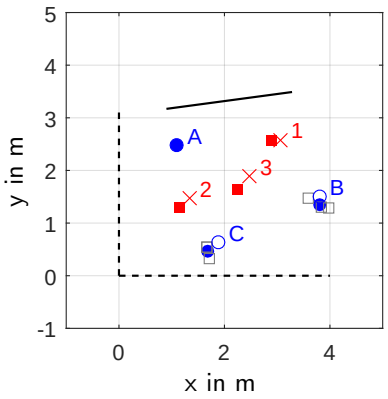
Figure 42a shows the results when using all three estimates found by each microphone and overlapping those resulting in a total of $N_{s,\text{est}} = N_i \times N_{\text{rep}} = 3 \times 6 = 18$ estimated points for the source positions and $N_{r,\text{est}} = N_{\text{rep}} = 6$ estimates for each microphone point. Figure 42b in contrast only uses the averaged points of each repetition to compute the error ellipses, resulting in $N_{s,\text{est}} = N_{r,\text{est}} = 6$ points used. For this

measurement, no significant difference in terms of the size of the error ellipses is observable when comparing the two figures.

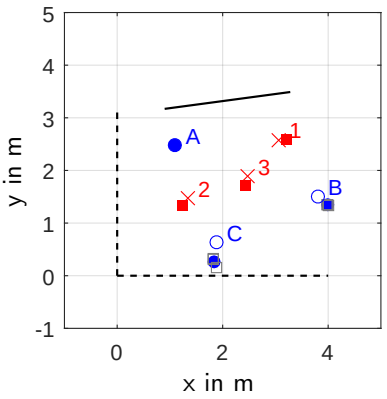
Results for the second measurements are shown in Figures 43 and 44 in the same way. Due to the larger number of microphones the use of all distinct source estimates results in $N_{s,\text{est}} = 8 \times 5 = 40$ (each repetition adds N_i source estimates, see Figure 43) and the use of the averaged source positions of each repetition in $N_{s,\text{est}} = 5$ (each repetition only contributes one source estimate, see Figure 44) source points for averaging to obtain the combined results and for computing the error ellipses. Comparing the results for the two case shows that the size of the error ellipses for the sources differ significantly, indicating that although the individual position estimates vary rather much, the averaged source positions of each repetition are nonetheless close together. This indicates that the results can be improved by using an increasing number of repetitions, i.e. more claps at each position and averaging to obtain the results increases the self-calibration accuracy. Furthermore it seems that sources which are surrounded by microphones exhibit a tendency to smaller as well as more uniform variances of the error distribution. Also the error ellipses of the microphone estimates (which are practically non existent in case of the second measurement) indicate that increasing the number of calibration-sources also increases the quality of the results.

The superposition results in numbers can be found in Table 6 and 7 for the first and second measurements, showing the MPE for each microphone and source position averaged over the repetitions. The results indicate that the self-calibration results when averaging over more than one repetition are actually much better than the MPEs of each separate repetition in Table 5 might let us assume, with the average MPE over the different microphones well below 10 *cm* and in the best case even below 1 *cm*.

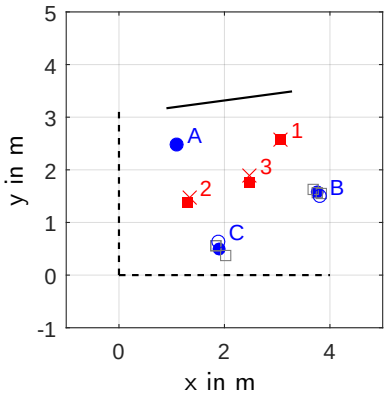
The orientation of each microphone array can be retrieved easily from the self-calibration results by finding the direction of 0° fixed by the array geometry such that all direct source DOAs point to the corresponding sources. The known orientation of each microphone array then allows the application of other algorithms and the use of the obtained results in combination with the estimated model.



(a) rep.1



(b) rep.2



(c) rep.3

Figure 40: Results of the source and microphone position estimates using the data form the first measurement (Appendix B.1), also showing the real position of the reflector.

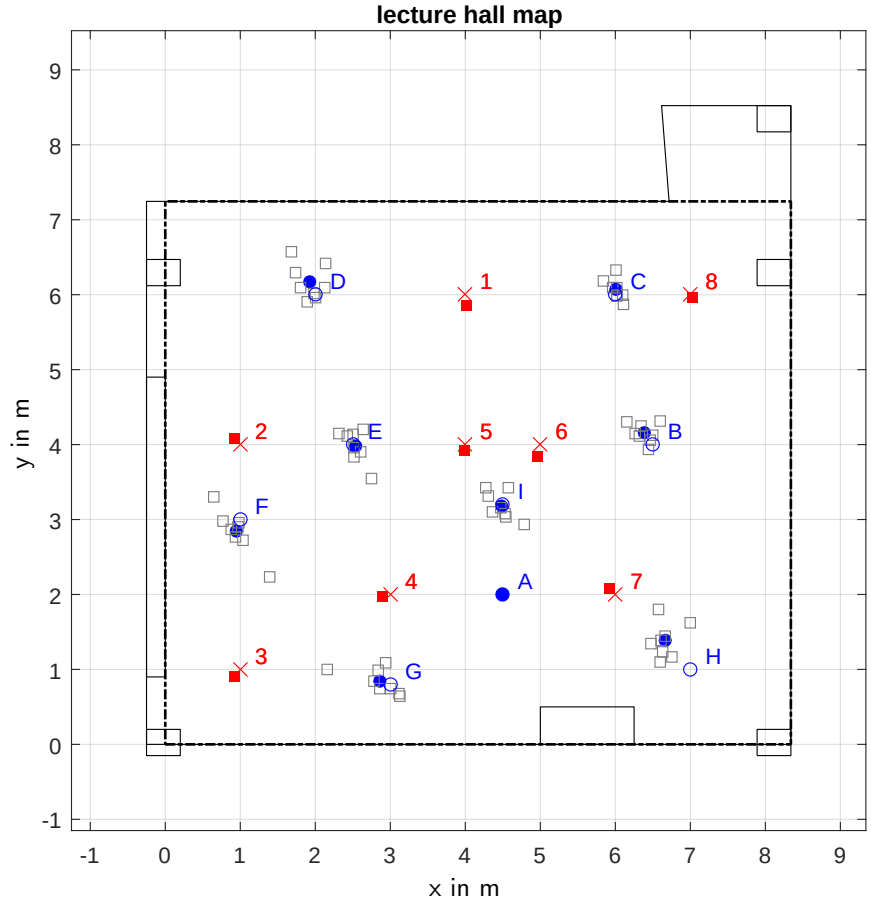
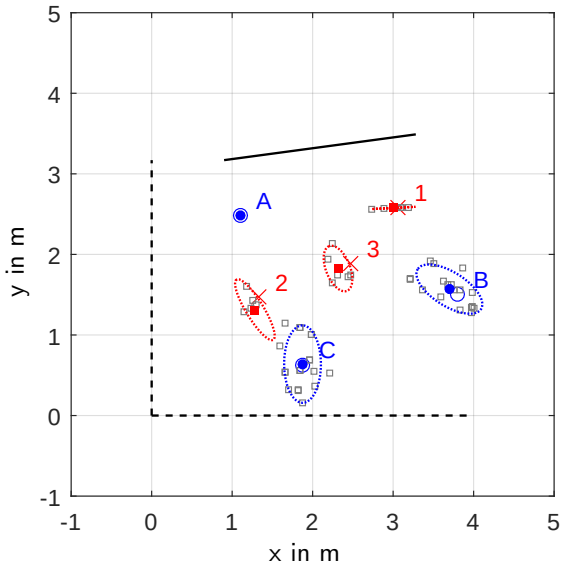
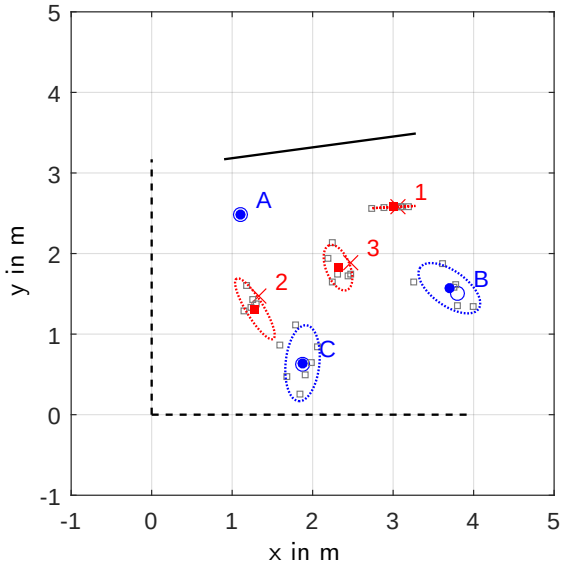


Figure 41: Results of the source and microphone position estimates. Important to notice is that it is known which $\hat{\mathbf{s}}_{ij}$ (□) belongs to which source when computing the averaging to find a final source estimate, so that overlapping positions clouds of $\hat{\mathbf{s}}_{ij}$ for different sources can be separated without the need of additional clustering.



(a) Error ellipses indicating the 70% confidence interval, computed using *all* estimated points.



(b) Error ellipses indicating the 70% confidence interval, computed from the *averaged* estimated points.

Figure 42: Combination of the microphone and source localization results for six repetitions from the first measurement. Position estimates averaged over the repetitions are indicated by ■ for the microphones and ● for the sources.

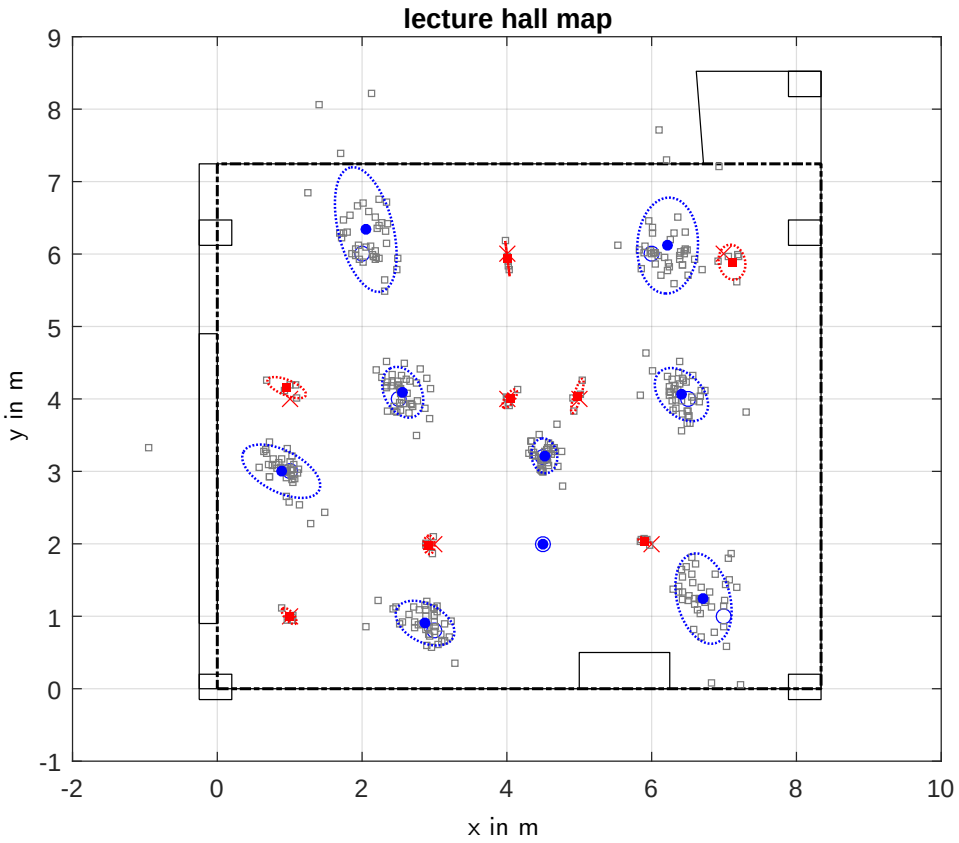


Figure 43: Combination of the self-calibration results for five repetitions, showing the error ellipses indicating the 70% confidence intervals, computed using *all points estimated at each repetition* for the source and microphone estimates in **blue** and **red** respectively. Position estimates averaged over the repetitions are indicated by ■ for the microphones and ● for the sources.

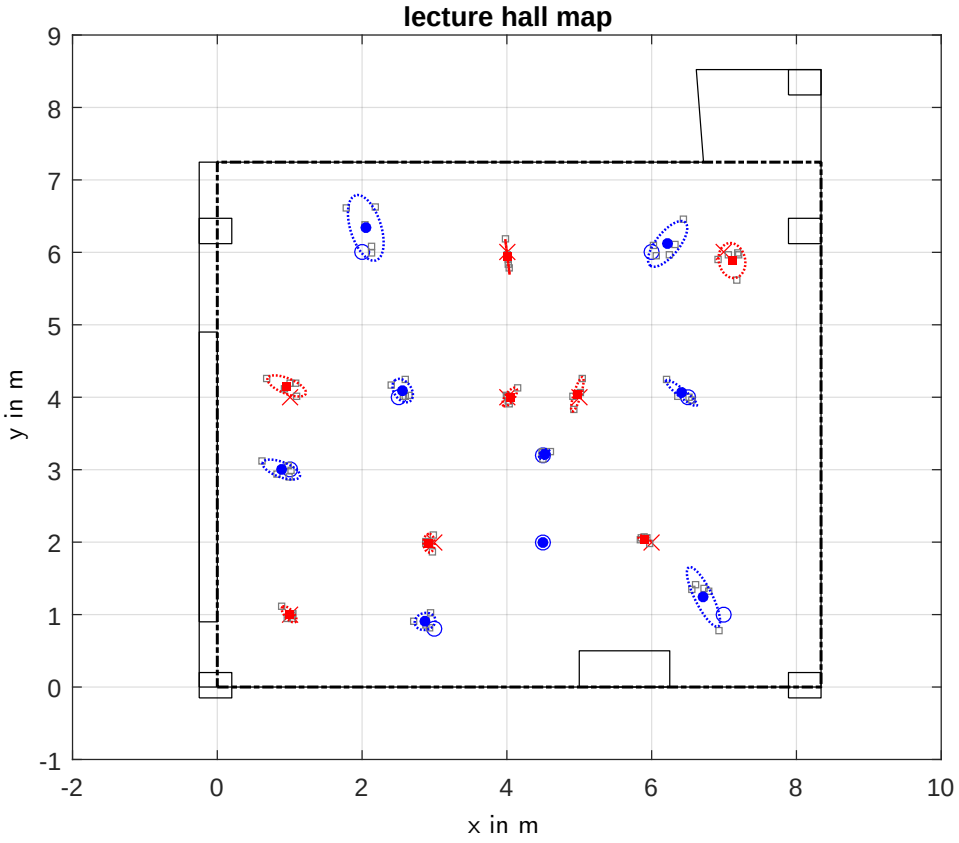


Figure 44: Combination of the self-calibration results for five repetitions, showing the error ellipses indicating the 70% confidence intervals, computed using the *estimated points averaged for each repetition* for the source and microphone estimates in **blue** and **red** respectively. Position estimates averaged over the repetitions are indicated by ■ for the microphones and ● for the sources.

Table 6: MPE and standard deviation for each microphone and source averaged over the results of six repetitions for the first measurement.

$\mathbf{r}_i$	$\varepsilon_{\mathbf{r},\text{ovr}}$ in m	$\sigma_{\mathbf{r},\text{ovr}}$ in m
1	0.0549	0.1236
2	0.1645	0.2074
3	0.1725	0.1725

$\mathbf{s}_j$	$\varepsilon_{\mathbf{s},\text{ovr}}$ in m	$\sigma_{\mathbf{s},\text{ovr}}$ in m
B	0.1228	0.1904
C	0.0105	0.1467

Table 7: MPE and standard deviation for each microphone and source averaged over the results of five repetitions for the second measurement.

$\mathbf{r}_i$	$\varepsilon_{\mathbf{r},\text{ovr}}$ in m	$\sigma_{\mathbf{r},\text{ovr}}$ in m
1	0.0651	0.0731
2	0.1617	0.1230
3	0.0090	0.0515
4	0.0755	0.0263
5	0.0405	0.0756
6	0.0475	0.0908
7	0.1071	0.0520
8	0.1570	0.1339
μ	**0.0829**	**0.0783**
σ	**0.0550**	**0.0366**

$\mathbf{s}_j$	$\varepsilon_{\mathbf{s},\text{ovr}}$ in m	$\sigma_{\mathbf{s},\text{ovr}}$ in m
B	0.1071	0.1431
C	0.2497	0.2259
D	0.3418	0.2536
E	0.1105	0.0769
F	0.1170	0.1544
G	0.1675	0.0884
H	0.3747	0.1379
I	0.0234	0.0408
μ	**0.1864**	**0.1401**
σ	**0.1239**	**0.0727**

4.1.3 Variation of Source and Microphone Numbers

In addition to the evaluation of the MPEs with the maximum number of microphones N_i and sources N_j is used the influence of varying the number of microphones and calibration sources is examined. For the first measurement results for fewer microphones and/or sources can be seen in Table 8 with the results at each (N_i, N_j)-combination evaluated for six repetitions. The resulting mean and standard deviations of the microphone and source MPEs are shown in Table 9. Overall, the results are better for more sources as well as more microphones, although low MPEs around $\sim 0.1\ m$ can be achieved for all source-microphone number combinations.

For the second measurement a more thorough examination can be performed due to the higher number of calibration sources and microphones that are available for variation. The averaged examples for the errors (mean MPE and standard deviation)

Table 8: Localization error for the estimated microphone and source positions from measurements one using different numbers of microphones N_i and sources N_j.

	Measurement 1					
	$N_i = 2$, $N_j = 3$		$N_i = 3$, $N_j = 2$		$N_i = 2$, $N_j = 2$	
rep.	$\varepsilon_{\mathbf{r}}$ in m	$\varepsilon_{\mathbf{s}}$ in m	$\varepsilon_{\mathbf{r}}$ in m	$\varepsilon_{\mathbf{s}}$ in m	$\varepsilon_{\mathbf{r}}$ in m	$\varepsilon_{\mathbf{s}}$ in m
1	0.2442	0.1754	0.2807	0.5300	0.8396	0.8220
2	0.1769	0.1723	0.1327	0.0428	0.0752	0.1029
3	0.0796	0.0603	0.3052	0.2467	0.6912	0.6194
4	0.0655	0.0538	0.7913	0.4022	1.0301	0.9363
5	0.6260	0.4542	0.8328	0.3189	1.6239	1.6623
6	0.2418	0.3414	0.1602	0.1140	0.4961	0.3836
μ	**0.2390**	**0.2096**	**0.4171**	**0.2758**	**0.7972**	**0.7544**
σ	**0.2045**	**0.1591**	**0.3133**	**0.1810**	**0.5220**	**0.5370**

of four clap repetitions for different combinations of microphones and sources numbers (all microphones and sources up to the respective number N_i or N_j are used) are visualized in Figure 45. Results in numerical values are given in Table 10, showing the mean values $\mu(\varepsilon_{\mathbf{r}})$ and $\mu(\varepsilon_{\mathbf{s}})$ as well as the standard deviations $\sigma(\varepsilon_{\mathbf{r}})$ and $\sigma(\varepsilon_{\mathbf{r}})$ computed over four repetitions for each (N_i, N_j)-variation. The lowest values are marked in **bold** letters and the three next closest values in *italics*. Examining the results show that the (N_i, N_j)-combinations resulting in small MPEs are located close to the bottom, i.e. for larger N_i, as well as near the right side, i.e. for larger N_j. The overall behaviour of the resulting microphone and source localization error is very similar to what could be expected from the results in Table 9, i.e. using more microphones and sources results in a lower MPE. Furthermore, increasing the number of microphones improves the results even when only two sources are used, while increasing the number of sources does not exhibit the same behaviour when using only two microphones. Using three or more microphones as well as increasing the source number again reduces the position errors. This might indicate that for a certain setup a minimum number of sources might be needed.

Figure 46 shows the evolution of the mean MPE of the microphone and source estimates averaged over five repetitions when using all sources ($N_j = 9$) while increasing the number of microphones from $N_i = 2$ to $N_i = 8$. An overall steady

Table 9: Mean and standard deviation of the microphone MPEs for different microphone numbers N_i and source numbers N_j.

$\mu(\varepsilon_{\mathbf{r}})$ in m	$N_j = 2$	$N_j = 3$
$N_i = 2$	0.7927	0.2390
$N_i = 3$	0.4171	0.2059

$\mu(\varepsilon_{\mathbf{s}})$ in m	$N_j = 2$	$N_j = 3$
$N_i = 2$	0.7544	0.2096
$N_i = 3$	0.2758	0.1792

$\sigma(\varepsilon_{\mathbf{r}})$ in m	$N_j = 2$	$N_j = 3$
$N_i = 2$	0.5520	0.2045
$N_i = 3$	0.3133	0.0944

$\sigma(\varepsilon_{\mathbf{s}})$ in m	$N_j = 2$	$N_j = 3$
$N_i = 2$	0.5370	0.1589
$N_i = 3$	0.1810	0.1078

improvement of the both position estimates can be observed[11].

Figure 47 shows the mean MPE of the estimated microphone and source positions when using all microphones ($N_i = 8$) and varying the number of sources from $N_j = 2$ to $N_j = 9$, averaging over five repetitions. As could be observed in Figure 45 both MPEs are already rather low when using $N_j = 2$ sources and $N_i = 8$ microphones. Small improvements concerning the standard deviation are still observable, although the mean microphone MPE is already below 0.25 m when using $N_j = 3$, which is rather low compared to the room size of roughly 8 by 7 meters.

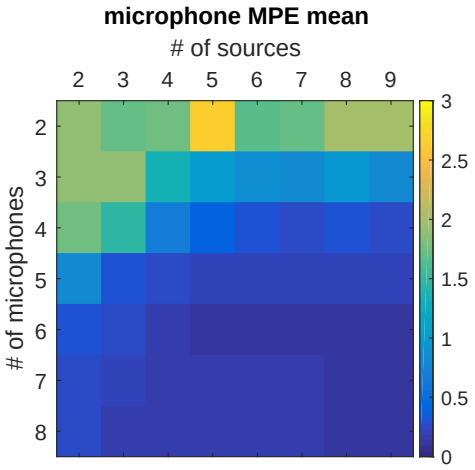
Figure 48 shows the mean MPEs of the estimated microphone and source positions when using two sources ($N_j = 2$) and a varying number of microphones N_i, again averaged over five repetitions. After initially rather high errors larger than 1 m, the mean MPE for the sources is steadily below 0.5 m when using $N_i = \{6, 7, 8\}$ microphones. The microphone mean MPE follows a similar shape, although the microphone estimates exhibit an initially slightly larger mean MPE.

It should be noted that varying the source and microphone numbers as presented here does not examine the spatial dependency of the MPEs on the source and microphone positions. In other words, when using N_i and N_j microphones and sources those are always all microphones/sources up to that number, e.g. for $N_i = 3$ and $N_j = 4$ the microphones and sources $\mathbf{r}_i$ with $i \in \{1, 2, 3\}$ and $\mathbf{s}_j$ with $j \in \{1, 2, 3, 4\} = \{A, B, C, D\}$. Examining permutations of the source and microphone combinations would be best performed with initial simulations followed by actual measurements to validate the simulations.

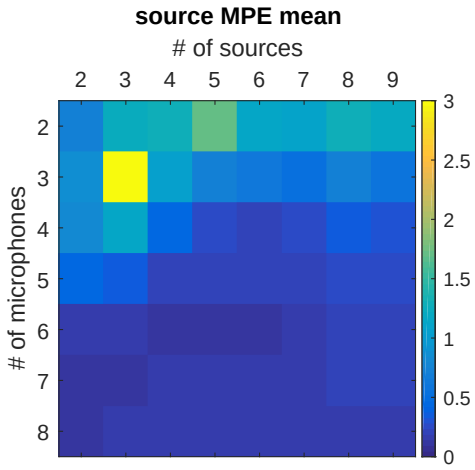
Furthermore this would result in extensive work concerned with finding good positions

[11] It should be noted that the position error of source A was not included when computing the results since it is always zero due to way the alignment was performed

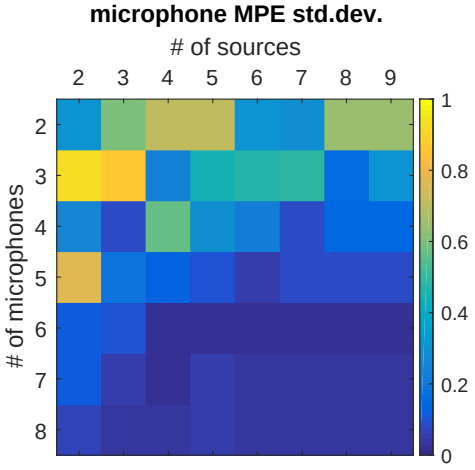
for sources and microphones based on the surrounding geometry and its acoustic properties (e.g. absorptive behaviour of the walls). The questions asked for actual measurements conducted in that context would furthermore be in the line of 'Do we need more/less sources/microphones in highly reverberant/exceptionally dry locations?' or 'Is there an ideal arrangement for sources and microphones resulting in minimal MPEs as well as needed calibration source and microphone numbers?', which would all be highly interesting but also well beyond the scope of this work.



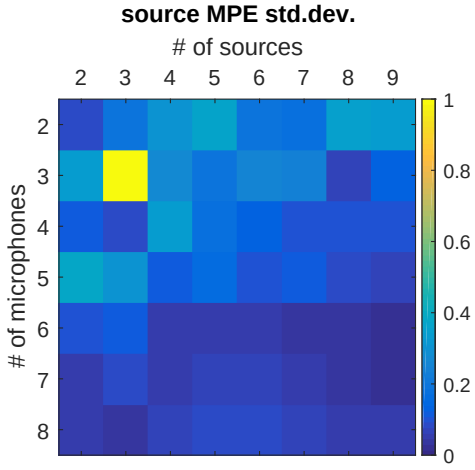
(a) $\mu(\varepsilon_{\mathbf{r}})$ in m



(b) $\mu(\varepsilon_{\mathbf{s}})$ in m



(c) $\sigma(\varepsilon_{\mathbf{r}})$ in m



(d) $\sigma(\varepsilon_{\mathbf{s}})$ in m

Figure 45: Mean MPE for different source and microphone numbers used for self-calibration. The x-axes show the number of sources, the y-axes the number of microphones used, with the MPE in m indicated by the colour of the field. The colors are bounded for better comparison as indicated by the colour bar (values higher than the corresponding maximum of the colorbar are set to the maximum color). The values are taken from Table 10.

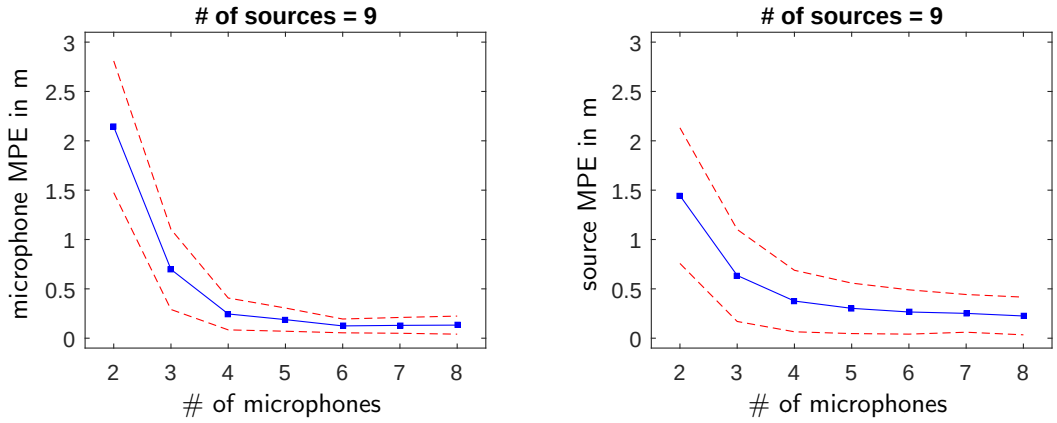
Table 10: Mean and standard deviation of the MPEs over four repetitions for all microphone and source number combinations N_i and N_j for the second measurement. The results are illustrated in as a color plot in Figure 45.

$\mu(\varepsilon_{\mathbf{r}})$ in m	$N_j = 2$	$N_j = 3$	$N_j = 4$	$N_j = 5$	$N_j = 6$	$N_j = 7$	$N_j = 8$	$N_j = 9$
$N_i = 2$	1.8897	1.7168	1.7732	2.6645	1.6840	1.6979	1.9992	1.9955
$N_i = 3$	1.9074	1.9167	1.3240	1.0300	0.8805	0.8256	0.9489	0.8308
$N_i = 4$	1.7531	1.4733	0.6944	0.3864	0.2815	0.2554	0.2827	0.2688
$N_i = 5$	0.8194	0.3245	0.2651	0.2229	0.1993	0.1929	0.1993	0.1968
$N_i = 6$	0.2871	0.2346	0.1651	0.1367	0.1269	*0.1255*	**0.1245**	*0.1274*
$N_i = 7$	0.2789	0.1986	0.1734	0.1721	0.1552	0.1418	0.1358	0.1343
$N_i = 8$	0.2443	0.1853	0.1723	0.1693	0.1488	0.1506	0.1357	*0.1337*

$\sigma(\varepsilon_{\mathbf{r}})$ in m	$N_j = 2$	$N_j = 3$	$N_j = 4$	$N_j = 5$	$N_j = 6$	$N_j = 7$	$N_j = 8$	$N_j = 9$
$N_i = 2$	0.3059	0.5997	0.7104	0.7168	0.3090	0.2966	0.6435	0.6427
$N_i = 3$	0.9234	0.8632	0.2457	0.4414	0.4834	0.4845	0.1602	0.3045
$N_i = 4$	0.2654	0.0864	0.5655	0.2916	0.2260	0.0895	0.1500	0.1548
$N_i = 5$	0.7803	0.2016	0.1342	0.1083	0.0586	0.0830	0.0849	0.0921
$N_i = 6$	0.1156	0.0940	*0.0257*	**0.0168**	0.0274	0.0282	0.0276	*0.0202*
$N_i = 7$	0.1140	0.0537	*0.0260*	0.0578	0.0434	0.0337	0.0316	0.0349
$N_i = 8$	0.0647	0.0366	0.0411	0.0569	0.0324	0.0328	0.0318	0.0329

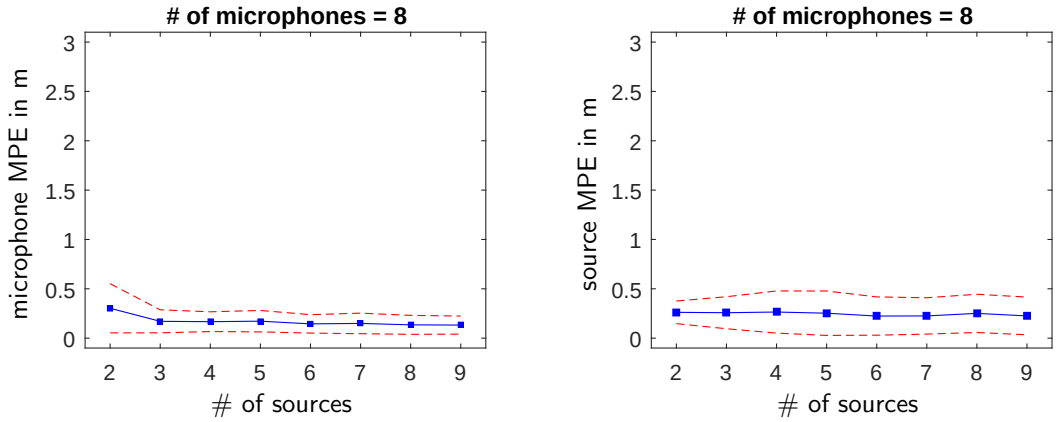
$\mu(\varepsilon_{\mathbf{s}})$ in m	$N_j = 2$	$N_j = 3$	$N_j = 4$	$N_j = 5$	$N_j = 6$	$N_j = 7$	$N_j = 8$	$N_j = 9$
$N_i = 2$	0.7251	1.2306	1.2998	1.7172	1.1531	1.0782	1.3012	1.2123
$N_i = 3$	0.8493	4.0162	1.0734	0.7481	0.6241	0.5554	0.7084	0.5769
$N_i = 4$	0.8105	1.1512	0.4386	0.2606	0.2042	0.2685	0.3403	0.2975
$N_i = 5$	0.4252	0.3613	0.2241	0.2036	0.1979	0.2228	0.2611	0.2415
$N_i = 6$	0.1596	0.1846	*0.1282*	*0.1269*	0.1370	0.1646	0.2046	0.1952
$N_i = 7$	*0.1370*	0.1396	0.1482	0.1766	0.1773	0.1726	0.2227	0.2004
$N_i = 8$	**0.1149**	0.1631	0.1591	0.1569	0.1468	0.1541	0.1811	0.1676

$\sigma(\varepsilon_{\mathbf{s}})$ in m	$N_j = 2$	$N_j = 3$	$N_j = 4$	$N_j = 5$	$N_j = 6$	$N_j = 7$	$N_j = 8$	$N_j = 9$
$N_i = 2$	0.0841	0.1904	0.3119	0.3668	0.1893	0.1781	0.3584	0.3324
$N_i = 3$	0.3430	5.5880	0.2714	0.1990	0.2628	0.2454	0.0686	0.1368
$N_i = 4$	0.1095	0.0826	0.3356	0.1768	0.1359	0.1009	0.0950	0.0961
$N_i = 5$	0.3858	0.3032	0.1132	0.1598	0.1020	0.1099	0.0864	0.0691
$N_i = 6$	0.1046	0.1249	0.0570	0.0508	0.0501	0.0453	*0.0345*	*0.0304*
$N_i = 7$	0.0560	0.0926	0.0558	0.0719	0.0718	0.0572	*0.0320*	**0.0297**
$N_i = 8$	0.0517	0.0435	0.0640	0.0928	0.0818	0.0670	0.0540	0.0486



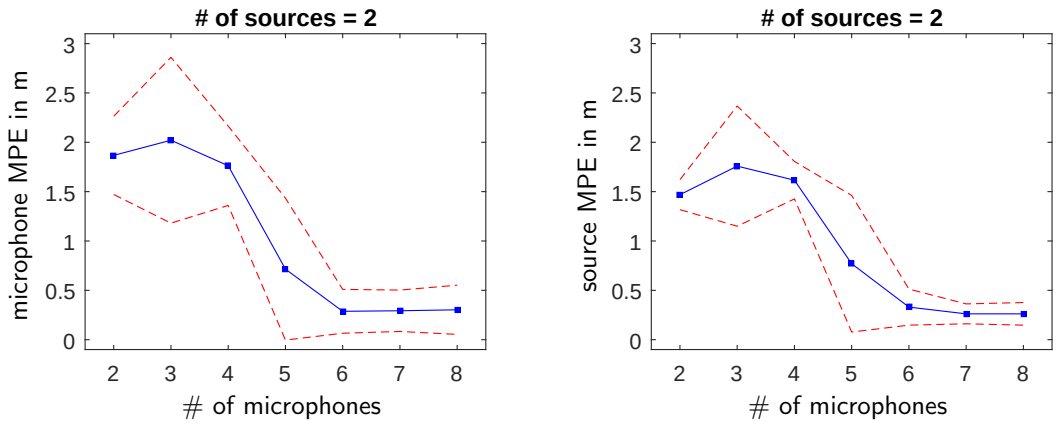
(a) $\mu(\varepsilon_{\mathbf{r}})$ for varying N_i **(b)** $\mu(\varepsilon_{\mathbf{s}})$ for varying N_i

Figure 46: Mean MPE and standard deviation for microphones and sources over five repetitions for varying numbers of microphones N_i. $N_j = 9$ sources are used.



(a) $\mu(\varepsilon_{\mathbf{r}})$ for varying N_j **(b)** $\mu(\varepsilon_{\mathbf{s}})$ for varying N_j

Figure 47: Mean MPE and standard deviation for microphones and sources over five repetitions for varying numbers of sources N_j. $N_i = 8$ sources are used.



(a) $\mu(\varepsilon_{\mathbf{r}})$ for varying N_i

(b) $\mu(\varepsilon_{\mathbf{s}})$ for varying N_i

Figure 48: Mean MPE and standard deviation for microphones and sources over five repetitions for varying numbers of microphones N_i. $N_j = 2$ sources are used.

4.2 Room Inference

4.2.1 Room Inference Error Measures

The evaluation of the room inference results (i.e. the localization of the reflectors) is performed on the Hough space parameters $\theta = (r, \vartheta)$, shown in Figure 2, as well as on the walls in vector form using $\mathbf{w}$ as the direction vector of the respective wall/reflector. The parameters are computed according to

$$\varepsilon_r = |r - \hat{r}| \tag{4.2.1}$$

$$\varepsilon_\vartheta = |\vartheta - \hat{\vartheta}| \tag{4.2.2}$$

$$\varepsilon_{\mathbf{w}} = \frac{\hat{\mathbf{w}}^T \mathbf{w}}{||\hat{\mathbf{w}}||_2 \cdot ||\mathbf{w}||_2}, \tag{4.2.3}$$

where r and $\hat{r}$ are the real and estimated orthogonal distance of the reflector to the origin, ϑ and $\hat{\vartheta}$ are the real and estimated perpendicular angles of the reflectors and $\mathbf{w}$ and $\hat{\mathbf{w}}$ are real and estimated direction vectors. The distance error ε_r is measured in meters, indicating the error in the distance of the estimated reflector, ε_ϑ the angular error in degrees. $\varepsilon_{\mathbf{w}}$ indicates the alignment error between the real and estimated reflector vectors, with values ranging from 0 to 1. Values closer to 1 indicate better alignment of the estimated and measured reflector.

From how the Hough parameters $\theta = (r, \vartheta)$ are defined it is obvious that for an accurate estimate both the distance and angular error have to be small. For a fixed setup (a microphone array, a calibration source and the estimated and corresponding real reflector), a low angle error $\varepsilon_\vartheta \approx 0°$ results in the estimated reflector being roughly parallel to the real one, with the distance error ε_r affecting the TDOA of the reflection caused by the estimated wall. A low distance error ε_r in turn indicates that the TDOA corresponding to the estimated wall is more or less correct, but that the DDOA of the reflection caused by the estimated reflector will be distorted.

4.2.2 Results when using all Microphones and Calibration Sources

The results are analysed for different repetitions of the first and second measurement, performed as described in Appendices B.1 and B.2. Important to notice is that the first measurement has only one reflector to estimate and that only two different estimation algorithms (PC projection and angle clustering) are examined. The second measurement in the more realistic environment contains four walls with four different estimation algorithms (PC projection, angle clustering, Hough transform

and rectangular fit) being examined. The reason for not using the Hough transform algorithm for the data from the first measurement was that there are usually only very few estimated reflection points available which proved difficult for this particular algorithm.

The results from the first measurement are summarized in Table 11 comparing the errors of the reflector estimates for PC projection and angle clustering. The results achieved by the angle clustering algorithm are better in terms of best and worst results as well as on average (six clap repetitions are examined), with the best result below 1 *cm* distance error and well below 1° angle error.

Table 11: Room inference results for the first measurement (Appendix B.1) containing a single reflector and using $N_i = 3$, $N_j = 3$.

	PC proj.			*angle clust.*		
rep.	ε_r in m	ε_ϑ in deg	$\varepsilon_\mathbf{w}$	ε_r in m	ε_ϑ in deg	$\varepsilon_\mathbf{w}$
1	0.3556	6.4017	0.9938	0.0608	0.5139	1
2	0.5225	2.0501	0.9994	0.2477	2.1652	0.9993
3	0.0624	1.2633	0.9998	0.0999	1.1340	0.9998
4	0.6804	9.0323	0.9876	0.4163	2.6863	0.9989
5	0.2780	4.5386	0.9969	0.1376	1.5780	0.9996
6	0.2643	4.3138	0.9972	0.0045	0.1640	1
μ	**0.3605**	**4.6**	**0.9958**	**0.1611**	**1.3736**	**0.9996**
σ	**0.2161**	**2.85**	**0.0046**	**0.1493**	**0.9645**	**4.3e-4**

The results for the second measurement are summarized in Tables 12, 13, 14 and 15, showing the errors for the respective walls termed north, east, south and west (shortened with n,e,s and w) for each implemented algorithm. For algorithms assuming a rectangular room also the average distance error μ_{ε_r} of all walls is computed as a measure of the overall quality of the estimate. For the Hough transform as well as the angle clustering, only results for the walls that could actually be estimated in some way are given.

The angle clustering yields better results in terms of best and worst results for the walls north, east and south, while the Hough transform performed error for the western wall. Furthermore, when using the angle clustering algorithm, estimates for all walls could be found while the Hough transform never detected the south wall, which was the one that received acoustic treatment resulting in very few reflection points detected on that wall at all repetitions.

The principal component projection and the rectangular fit (without the additional

rotation parameter) suffer mostly from an angular error of the estimated walls as they are both dependent on the principal components of the estimated reflection points being parallel to the actual walls. As can be deduced from the results of *rep.*1 in Tables 14 and 15 this assumption can still lead to good results. Since the angle errors for both are identical, the only difference between these two algorithms are the distances of the estimated walls, where the rectangular fit usually performs better, when comparing the best and worst case results for the distance error as well as on average.

Including the additional rotation parameter to the rectangular fit algorithm (as described in Section 3.3.6) eliminates the influence of the principal components of the estimated reflection points. A comparison of the rectangular fit algorithm with and without the additional rotation parameter as well as when using a higher threshold for the weights of the reflection points can be seen in Table 16. The additional rotation significantly reduces the angular error (i.e. the rotation error of the room) as well as the average distance error of the estimated room. The numerical results presented in Table 16 are shown in Figure 49 for visual comparison, showing that the misalignment of the real and estimated walls is reduced.

Table 12: Room inference results for angle clustering.

AC	ε_r in m				ε_ϑ in deg				$\varepsilon_{\mathbf{w}}$			
rep.	n	e	s	w	n	e	s	w	n	e	s	w
1	0.31	0.18	-	0.69	4.2	3.3	-	5.1	0.9971	0.9983	-	0.9960
2	0.20	0.03	-	1.65	4.0	4.1	-	15.8	0.9976	0.9974	-	0.9624
3	0.10	-	-	-	0.3	-	-	-	1	-	-	-
4	0.75	0.42	-	-	7.5	7.1	-	-	0.9915	0.9923	-	-
5	0.01	-	0.32	-	0.1	-	2.3	-	0.9999	-	0.9992	-

Table 13: Room inference results for Hough transform.

HT	ε_r in m				ε_ϑ in deg				$\varepsilon_{\mathbf{w}}$			
rep.	n	e	s	w	n	e	s	w	n	e	s	w
1	0.28	0.28	-	0.23	1.3	2.7	-	1.7	0.9998	0.9989	-	0.9995
2	0.64	0.45	-	0.31	9.3	2.1	-	1.6	0.9866	0.9994	-	0.9996
3	0.24	0.32	-	-	1.4	12.9	-	-	0.9997	0.9744	-	-
4	1.06	0.46	-	0.77	8.4	7.2	-	15.6	0.9891	0.9922	-	0.9632
5	0.29	0.08	-	0.65	3.9	5.4	-	15.3	0.9979	0.9956	-	0.9646

Table 14: Room inference results for principal component projection.

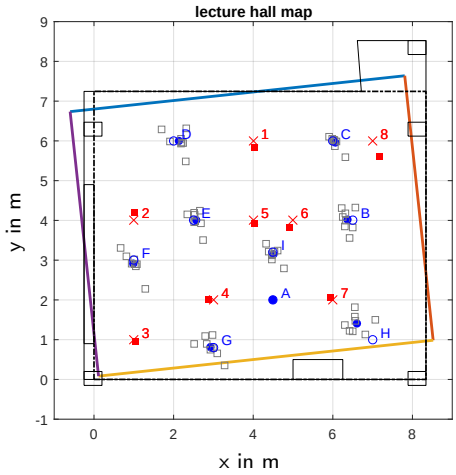
PC	ε_r in m					ε_ϑ in deg	$\varepsilon_{\mathbf{w}}$
rep.	n	e	s	w	μ_{ε_r}		
1	0.23	0.04	0.25	0.27	**0.20**	0.3	1
2	0.50	0.40	0.70	0.96	**0.64**	8.7	0.9885
3	1.12	1.71	0.88	1.53	**1.31**	14.9	0.9665
4	0.34	0.78	0.98	0.28	**0.60**	12.0	0.9780
5	2.17	0.74	0.57	0.23	**0.93**	12.7	0.9757
μ					**0.74**	**9.72**	**0.98**

Table 15: Room inference results for the rectangular fit algorithm without the additional rotation parameter.

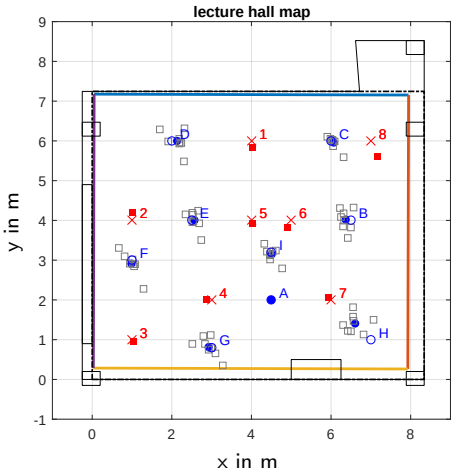
RF	ε_r in m					ε_ϑ in deg	$\varepsilon_{\mathbf{w}}$
rep.	n	e	s	w	μ_{ε_r}		
1	0.01	0.14	0.06	0.11	**0.08**	0.3	1
2	0.56	0.63	0.10	0.78	**0.52**	8.7	0.9885
3	1.15	1.21	1.23	1.27	**1.22**	14.9	0.9665
4	0.57	0.93	0.68	1.27	**0.86**	12.0	0.9780
5	0.61	0.68	0.01	0.71	**0.50**	12.7	0.9757
μ					**0.64**	**9.72**	**0.98**

Table 16: Comparison of room inference result for the rectangular fit without and with the additional rotation (marked by ↻), with a higher threshold used on the reflection points.

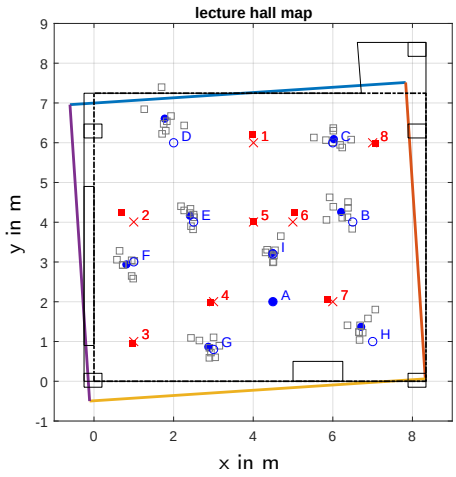
RF	ε_r in m					ε_ϑ in deg	$\varepsilon_{\mathbf{w}}$
rep.	n	e	s	w	μ_{ε_r}		
1	0.48	0.24	0.07	0.12	**0.23**	6.13	0.9943
1 ↻	0.07	0.41	0.29	0.03	**0.20**	0.16	1
2	0.26	0.03	0.49	0.14	**0.23**	3.79	0.9978
2 ↻	0.08	0.26	0.25	0.28	**0.22**	0.77	1
3	1.58	3.27	0.88	3.51	**2.31**	40.22	0.7636
3 ↻	0.02	0.52	0.52	0.10	**0.29**	0.23	1



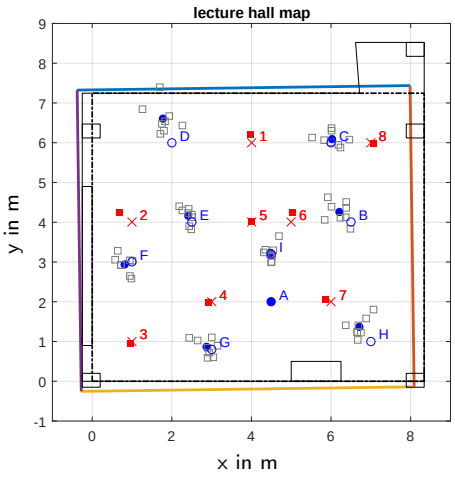
(a) no rotation, rep.1



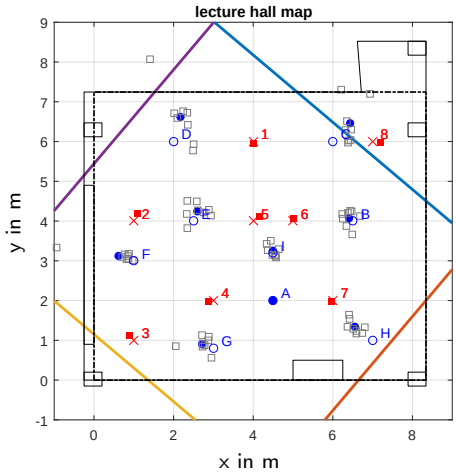
(b) rotation, rep.1



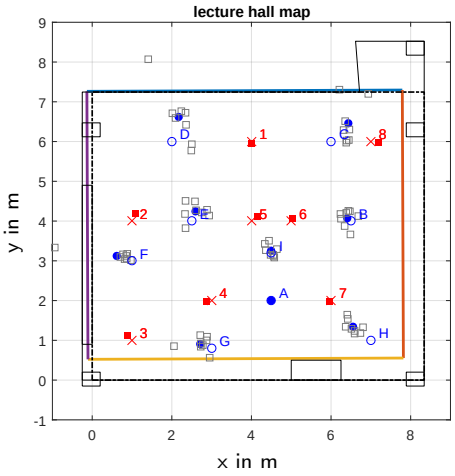
(c) no rotation, rep.2



(d) rotation, rep.2

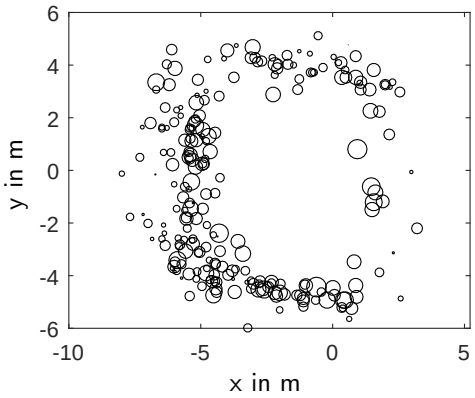


(e) no rotation, rep.3

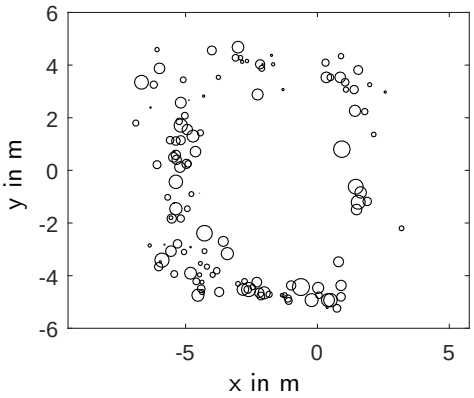


(f) rotation, rep.3

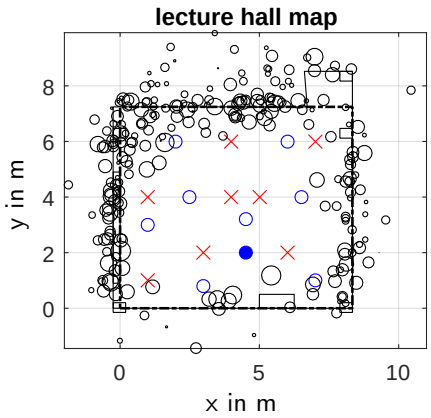
Figure 49: Room inference results for rectangular fit algorithm without (left column) and with (right column) the additional rotation parameter. Both algorithms use a higher threshold on the estimated reflection points.



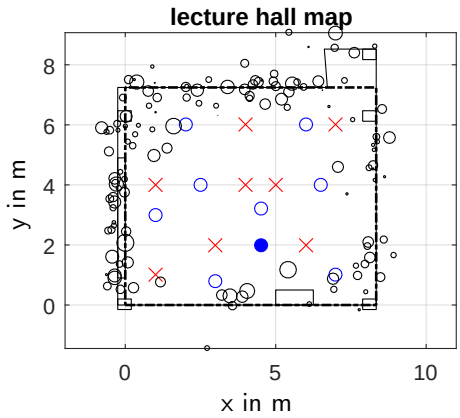
(a) low threshold, $N_{\text{refl}} = 238$



(b) high threshold, $N_{\text{refl}} = 126$



(c) low threshold, $N_{\text{refl}} = 238$



(d) high threshold, $N_{\text{refl}} = 126$

Figure 50: Comparison of estimated reflection points (○) for increased weight threshold, indicating the number of reflection points N_{refl} for the first repetition. The size of the circle reflects the weight of the reflection point.

The influence of low and high thresholds for the reflection points used for the room inference task can be seen in Figure 50, with the number of reflection points N_{refl} used in each case indicated. For both the low threshold resulting in $N_{\text{refl}} = 238$ as well as for the high threshold with $N_{\text{refl}} = 126$ the shape of the room can be seen clearly when looking at the unaligned results[12], though the low threshold still includes some far off points that might cause problems. The size of the circles depicting the reflection points indicate the underlying weight.

[12]The alignment corresponds to a clockwise rotation of roughly 90° and a shift of origin.

4.2.3 Variations of Source and Microphone Numbers

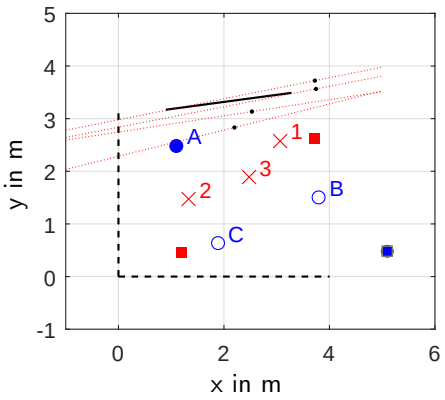
Thinking about the computation the reflection points, it is obvious that the quality of the estimated reflector positions will somehow be linked to the quality of the estimated source and microphone positions. Results for the estimation of the single reflector with varying numbers of calibration sources and microphones for data from the first measurement (described in Appendix B.1) are presented in Figure 51. The corresponding numerical values can be found in Table 17.

When the source and microphone positions are localized with low MPE (repetition 2 in Table 17 and Figures 51c and 51d) the errors for distance, angle and alignment are low as well. In spite of the inaccurate position estimates in repetitions 1 and 3 the results of the estimated reflectors remain rather close to the real positions. This is due to the fact that the DOA of the reflection point is anchored to the DOA of the direct sound which can in turn only move closer to or farther away from the microphones by increasing or decreasing the TOFs in the self-calibration problem. A direct result of this is that the angles between the estimated points remain very close to those between the real points, thus only resulting in some form of 'scaling' of the overall model. The problems are therefore most likely introduced by the TOA estimation stage.

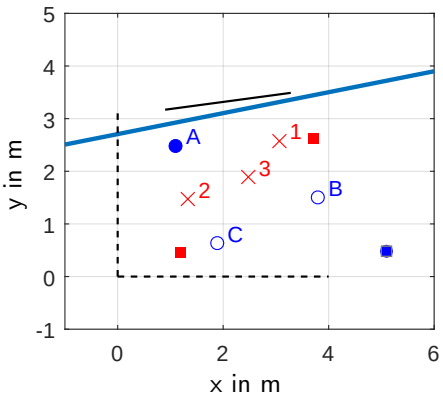
Owing to the few reflection points that are estimated, no clustering algorithms are used. The reflector location and angle are instead found by averaging over positions and angles of the two reflection points with the largest weights.

Table 17: Reflector localization results when using a reduced number of sources and microphones on data of the first measurement.

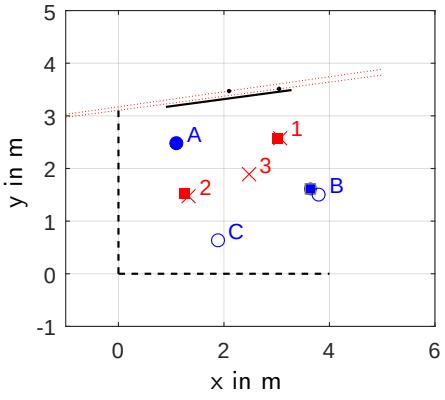
	$N_i = 2,\ N_j = 2$		
rep.	ε_r in m	ε_ϑ in deg	$\varepsilon_{\mathbf{w}}$
1	0.368	3.60	0.998
2	0.086	0.21	1
3	0.505	4.39	0.9971



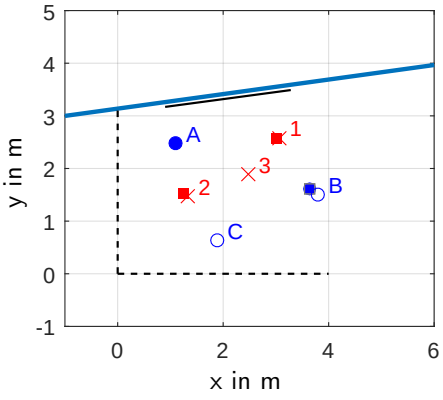
(a) all reflection points, rep.1



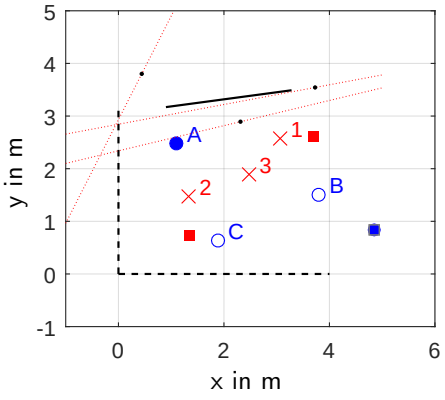
(b) averaged reflector, rep.1



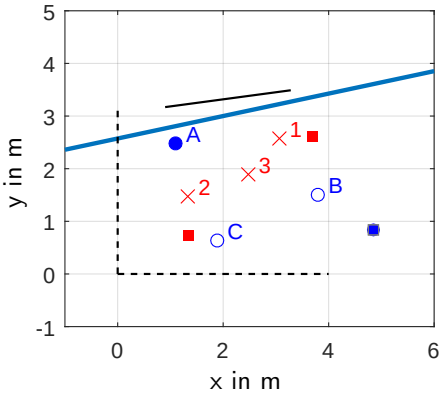
(c) all reflection points, rep.2



(d) averaged reflector, rep.2



(e) all reflection points, rep.3



(f) averaged reflector, rep.3

Figure 51: Source and microphone position estimates with estimated reflection points (left) and averaged reflector (right) using the minimum number of sources and microphones $N_i = 2$ and $N_j = 2$.

4.3 Room Height

The results for estimating the room height for measurements performed in the lecture hall (see Appendix B.2) can be seen in Figure 52 as well as in Table 18 as numerical results. The original height of the room is $h_{\text{real}} = 3.08\ m$ according to the construction plan. Table 18 shows the estimated height h_{est} which is computed as

$$h_{\text{est}} = \hat{h}_c + \hat{h}_f \tag{4.3.1}$$

and the error ε_h as the absolute difference between the real and estimated heights computed as

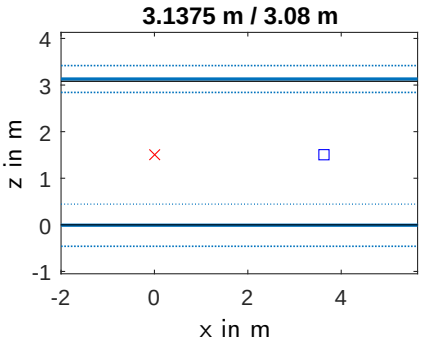
$$\varepsilon_h = |h_{\text{real}} - h_{\text{est}}|. \tag{4.3.2}$$

Figure 52 shows the locations of the estimated floor and ceiling relative to the plane of the sources and microphones for different repetitions, visualizing how the separate errors of floor and ceiling distance combine into the overall error. The standard deviation bounds are shown as dashed lines above and below floor and ceiling estimates.

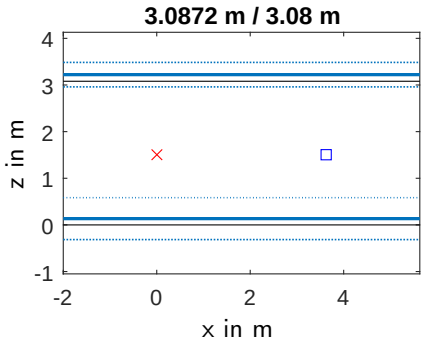
In the best case an overall height error below 1 *cm* is achieved, with the overall height error often partitioned more or less equally into the estimated distance to floor and ceiling as indicated by the results shown in Figures 52. Overall, the average height error of 13 *cm* is low compared to the room height. When averaging over the distinct height estimates of each repetition, an even lower error of 5.18 *cm* (comparing the averaged height of 3.0282 *m* with the real one of 3.08 *m*) is achieved, as already observed when combining the self-calibration results over more repetitions.

Table 18: Height estimation for different clap repetitions.

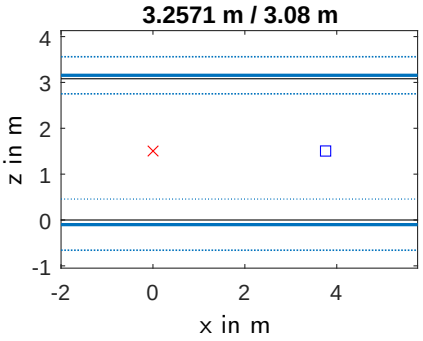
rep.	h_{real} in m	h_{est} in m	ε_h in m
1	3.08	3.1375	0.0575
2	3.08	3.2571	0.1771
3	3.08	2.8700	0.2100
4	3.08	3.0872	0.0072
5	3.08	2.8789	0.2011
6	3.08	2.9386	0.1414
μ		**3.0282**	**0.1324**
σ		**0.1570**	**0.0826**



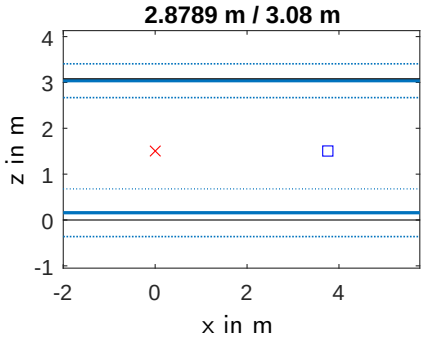
(a) estimated height, rep.1



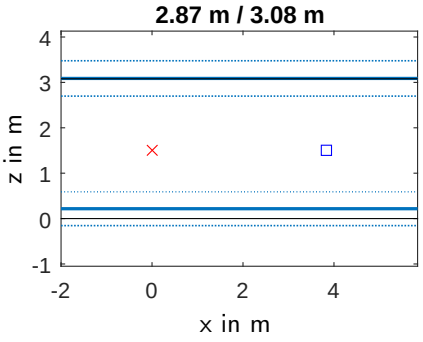
(b) estimated height, rep.4



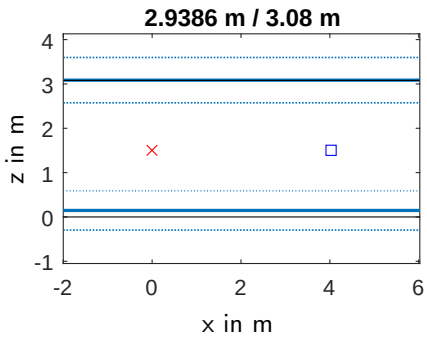
(c) estimated height, rep.2



(d) estimated height, rep.5



(e) estimated height, rep.3



(f) estimated height, rep.6

Figure 52: Floor and ceiling estimates (solid blue lines) with the standard deviation shown (dashed blue lines). An exemplary source and microphone are indicated by the usual symbols at the height of 1.5 m where they were positioned for the measurements. The real floor and ceiling are indicated by solid black lines at 0 and 3.08 m.

5 Conclusion

This work developed a complete set of tools for the self-calibration of an array of distributed first order microphones and, based on these results, for performing room inference. It consists of a parameter estimation stage followed by a scene reconstruction stage performing the self-calibration and the room inference.

The estimated parameters are the TOAs (*time of arrivals*) and DOAs (*directions of arrival*) of direct sound and reflections, both estimated in the frequency domain. The proposed instantaneous DOA estimator is an extension of an existing algorithm by Politis et al. [PDMP15], that allows DOA estimates at frequencies above the aliasing frequency of the used microphone arrays, which is extended by subspace processing. The proposed TOA estimation algorithm also uses a separation into signal and noise subspace to find TOAs of direct sound and reflections as time instances with large signal energy (i.e. a large signal eigenvalue).

Using the found TOAs and DOAs, the inter-microphone *time differences of arrival* (TDOAs) as well as the *direction differences of arrival* (DDOAs) of different direct sounds are fused together in the proposed self-calibration algorithm to find the positions of the microphones and calibration sources up to a arbitrary rotation.

The TDOA and DDOA pairs that were assigned to reflections are then used to estimate reflection points (assuming known source and microphone locations) on which four algorithms perform the actual inference. Two of these algorithms allow the detection of arbitrary linear reflectors with the other two specialized on rectangular geometries.

A time based method for localizing the floor and ceiling based on the self-calibration results is implemented by finding ellipses that fulfill the TDOA constraints of reflections assumed to stem from floor and ceiling (a method already popularly used in literature).

The developed algorithms are easy to maintain, exchange or update. The self-calibration stage only uses the direct sound TOAs and DOAs and the room inference only relies on known source and microphone locations and TOA and DOA pairs of detected reflections (i.e. the results from the self-calibration and parameter estimation stages). How these TOA and DOA pairs are acquired has no effect on the algorithms. All weights used in the algorithms are optional, though suitable weights

should be possible to find in other TOA and DOA estimation algorithms as well.

Most importantly it is shown that the developed algorithms are already fit for use on real world data, with all results acquired using actual measurement data taken from two measurements. These are conducted in an acoustically treated measurement room as well as in a medium sized lecture hall with basic acoustic treatment. As calibration source manually performed claps are used, allowing fast and simple repetitions which can be performed easily by a single person. Neither prior knowledge (apart from the type of microphone array used and assumptions on the speed of sound) nor manual position measurements are needed. The claps used for self-calibration and geometry inference are performed consecutively and then fused together when estimating the TOAs and DOAs (i.e. more claps can be performed until the estimated positions are accurate enough).

In the measurement room, a mean position error of 20 *cm* for the microphones and 18 *cm* for the calibration sources is achieved on average, with best case results of 8.3 *cm* for the microphones and 7.4 *cm* for the calibration sources. The distance and orientation errors for the single reflector are 16 *cm* and 1.4° on average, with the best case estimates resulting in errors of 0.45 *cm* and 0.16°.

Similar results are produced in the lecture hall with a mean position error of 17 *cm* for the microphones and 21 *cm* for the calibration sources on average, and 9 *cm* and 15 *cm* for microphones and calibration sources respectively in the best case. It is furthermore shown that the average error of the microphones can be reduced by performing consecutive measurements and combining the results, resulting in a mean position error of 8 *cm* for the microphones and 18 *cm* for the sources. The room inference in the lecture hall yields good results as well, with average distance errors of the four walls of roughly 20 *cm* and an average angular error of the estimated rectangular room below 0.5°. When estimating the height of the lecture hall, an average error of 13 *cm* is achieved. Averaging the results over more repetitions also improves the estimated height, resulting in an error of roughly 5 *cm*.

5.1 Applications and Future Work

The developed set of algorithms is ready for use with existing microphone arrays, such as for example the WiLMA project described by Schörkhuber et al. in [SZZ14] which uses the same microphone arrays. Moreover, adaptions for other microphone arrays than the ones used in this work are straight forward by either simple adaptions of the DOA estimation stage, or by exchanging parts with already existing algorithms.

Work that is still to be done would be to create a program that can be used in a stand-alone way, as well as to reduce the computation time by optimizing the algorithms.

Another very interesting topic which was hardly scratched in this work would be to examine the interactions of different room geometries exhibiting different absorptive properties with the locations of the calibration sources and microphones. This would immensely help by giving insight into which calibration source positions are futile or how many sources are actually needed for calibrating a given microphone array in a certain environment, possibly saving much time.

Also, the proposed DOA estimator leaves room for improvement in terms of the quality of the estimated elevation DOA, although this might be countered by either rotating some microphone arrays by 90° from the horizontal plane (i.e. towards floor or ceiling) or by performing two consecutive measurements and rotating all microphones in the second measurement (without changing the positions).

Further examinations of the results by using an auralization of an estimated scene model for direct comparison with the real scene, or using the estimated models to warp the analyzed room into another one by changing estimated parameters might be interesting in the context of psychoacoustic experiments. This would be particularly interesting with respect to possible applicability in virtual acoustics.

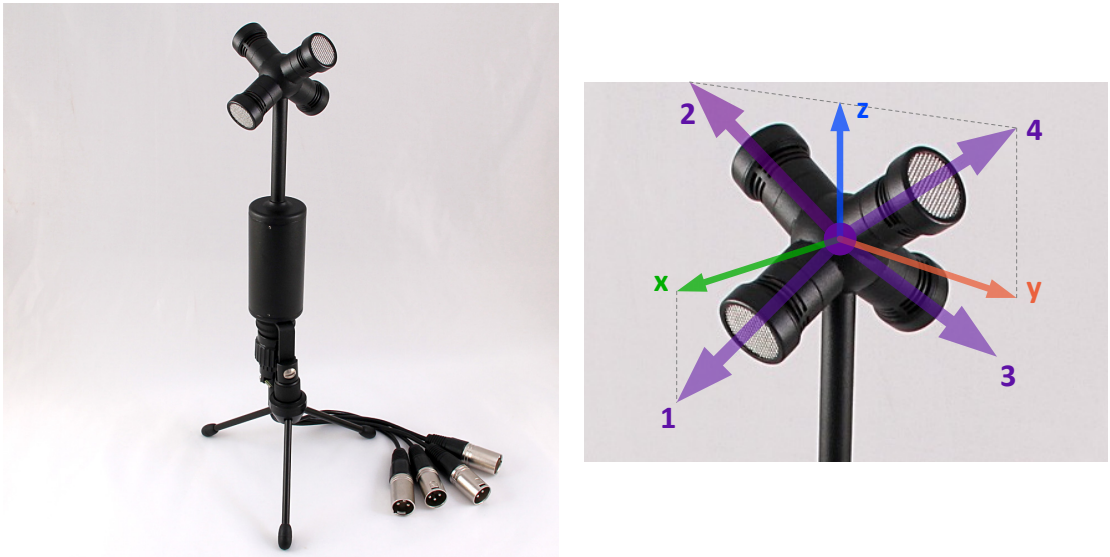
Particularly interesting would be the use of an estimated model to examine possible improvements on beamforming algorithms or source signal extraction, in the field of ambient assisted living or in live recordings of various types.

A Oktava Ambient 4D

The microphone arrays used are the *Oktava Ambient 4D*, which are B-Format microphones with the advantage that they record in A-Format. A picture of a microphone array can be seen in Figure 53a. The coordinates of the microphone capsules in Cartesian coordinates with respect to the juncture of the capsules are given as

$$\mathbf{N}_K = \begin{pmatrix} | & | & | & | \\ \boldsymbol{\nu}_1 & \boldsymbol{\nu}_2 & \boldsymbol{\nu}_3 & \boldsymbol{\nu}_4 \\ | & | & | & | \end{pmatrix} = \begin{pmatrix} 0.0286 & 0 & -0.0286 & 0 \\ 0 & -0.0286 & 0 & 0.0286 \\ -0.0202 & 0.0202 & -0.0202 & 0.0202 \end{pmatrix}, \quad \text{(A.0.1)}$$

where each column corresponds to one look direction of a capsule. The numbering as well as the local coordinate system used are depicted in Figure 53b. The values in Equation A.0.1 are adopted form the work by Hack [Hac15].



(a) Oktava Ambient 4D **(b)** Capsule vectors for each capsule.

Figure 53: Pictures of an Oktava microphone array, taken from [Okt]. The overlays in Figure 53b are added for clarification of Equation A.0.1.

B Measurements

Two measurements in different environments are performed for this work, the first one conducted in a room designed for acoustic measurements and the second in a lecture room with basic acoustic treatment. The first measurement is used to examine the conditions to be dealt with and initial algorithm development as well as for assessment of algorithm performance later on. The second measurement is then used for evaluation under more realistic conditions. The following sections describe the measurements, detailing the execution and giving a ground truth map of each measurement setup. The speed of sound was assumed to be $c = 340\frac{m}{s}$ throughout this work.

B.1 First Measurement

The first measurement is rather simple, using a single reflective surface in an otherwise reflection free room, three microphones and three different source positions, with the source signal being a hand produced clap. As reflective surface, a stack of tables set on their sides is used, stretching from the floor almost to the ceiling. To get more data out of each calibration source position different types of claps are used at each position, termed *broadband*, *low pass* and *high pass*, indicating the acoustic impression while clapping and are performed consecutively at each location.

A map of the measurement setup can be seen in Figure 54 with the reflective surface shown as a solid black line and the walls of the measurement room (assumed to be absorptive) as a dashed black line. The microphone positions are marked by a red $\times$ indexed with $\{1, 2, 3\}$, the source positions by a blue $\circ$ indexed by {A,B,C}. All microphones and sources are one the same level of 1.5 m, whereas the reflective surface stretched approximately 1 m above and 1.5 m below that level. It should be noted that due to the fact that the source signals are produced by manual clapping, small positioning errors might have been caused by the performer, although the chosen positions were marked on the floor before the measurement.

The microphones used are *Oktava 4D* microphones (described in Appendix A). A detailed view of waveforms captured by the microphone arrays can be seen in Figure 55, showing the discrete capsule signals with the direct sound and a reflection thereafter observable.

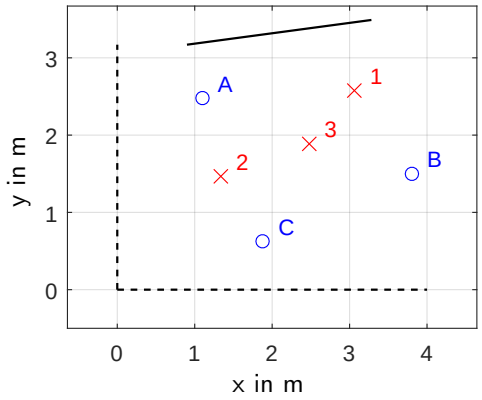
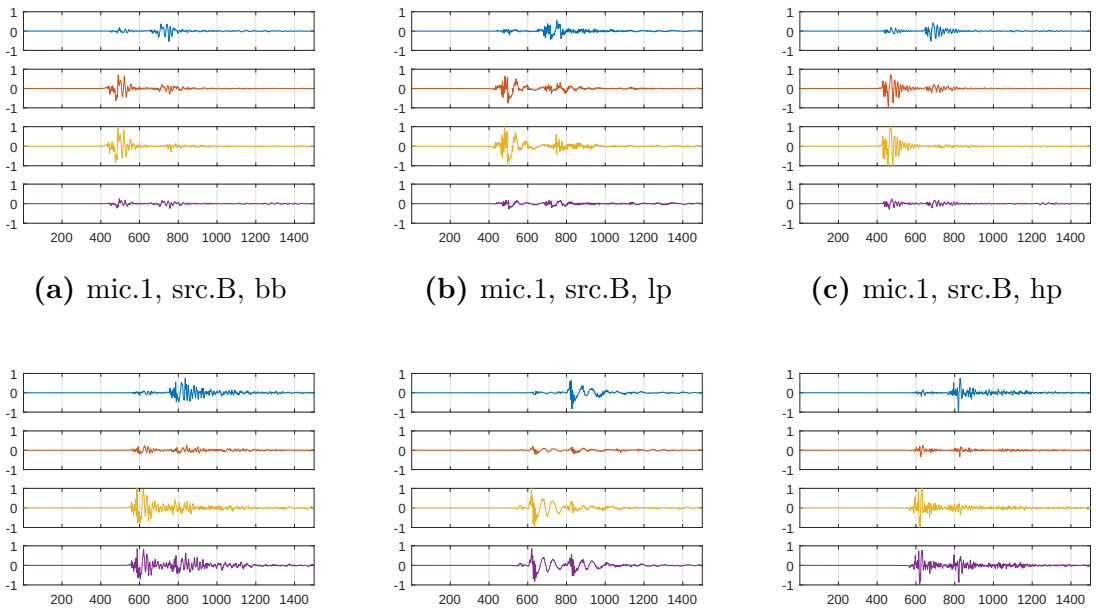


Figure 54: Room Model of the first measurement with the reflective surface indicated. The black dashed line corresponds to the walls of the measurement room assumed to be absorptive, the solid black line to the reflective surface. Microphones are marked by × and **1-3** and sources by ○ and **A-C**.



(a) mic.1, src.B, bb **(b)** mic.1, src.B, lp **(c)** mic.1, src.B, hp

(d) mic.1, src.C, bb **(e)** mic.1, src.C, lp **(f)** mic.1, src.C, hp

Figure 55: Microphone signals for each capsule of the first measurement, shown to visualize the possibility to detect the reflection as well as the amplitude differences at different capsules (amplitudes of each microphone array signal are normalized to 1, the x-axis shows the time in *samples*). The clap types broadband (bb), low pass (lp) and high pass (hp) and microphone/source number are indicated.

B.2 Second Measurement

The second measurement is performed in a lecture hall to acquire data from a more realistic surrounding. Still, to create an environment with reasonable reflections the tables and chairs are stacked and moved to the side to create room for positioning the microphones and the clapping person.

The lecture hall is a medium size room that received basic acoustic treatment (acoustic foam on the back wall) with a floor area of 59.31 m^2 and a room height of 3.08 m. The room map in Figure 56 shows the microphone and calibration source positions, as well as the location from where the panorama shot shown in Figure 57 was taken (upper left corner at coordinates $(1, 6.5)$ m approximately). The second measurement is performed identical to the first one in terms of clap types, only with twice the number of claps performed at each location.

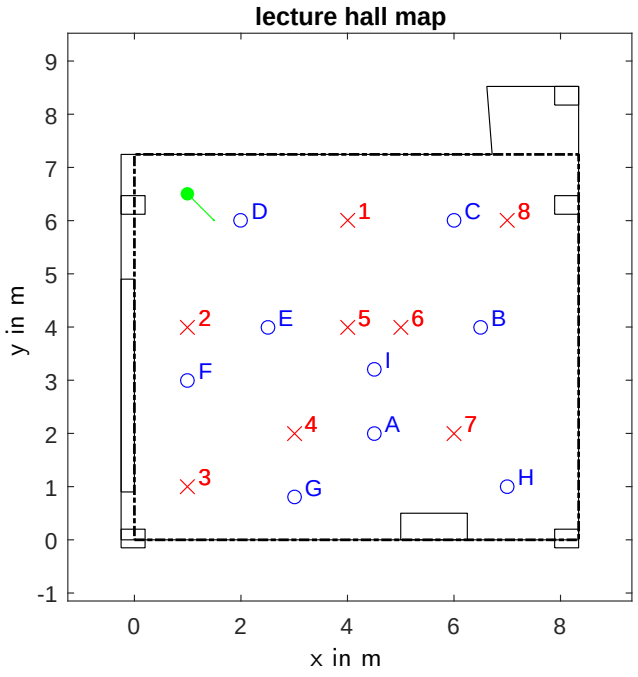


Figure 56: Floor map for the second measurement. The positions of microphones **1-8** are marked by ×, the positions of the calibration sources where the claps are performed with ○ and indexed by **A-I**. The location and look direction from where the panorama picture shown in Figure 57 is taken is marked in green in the upper left corner.

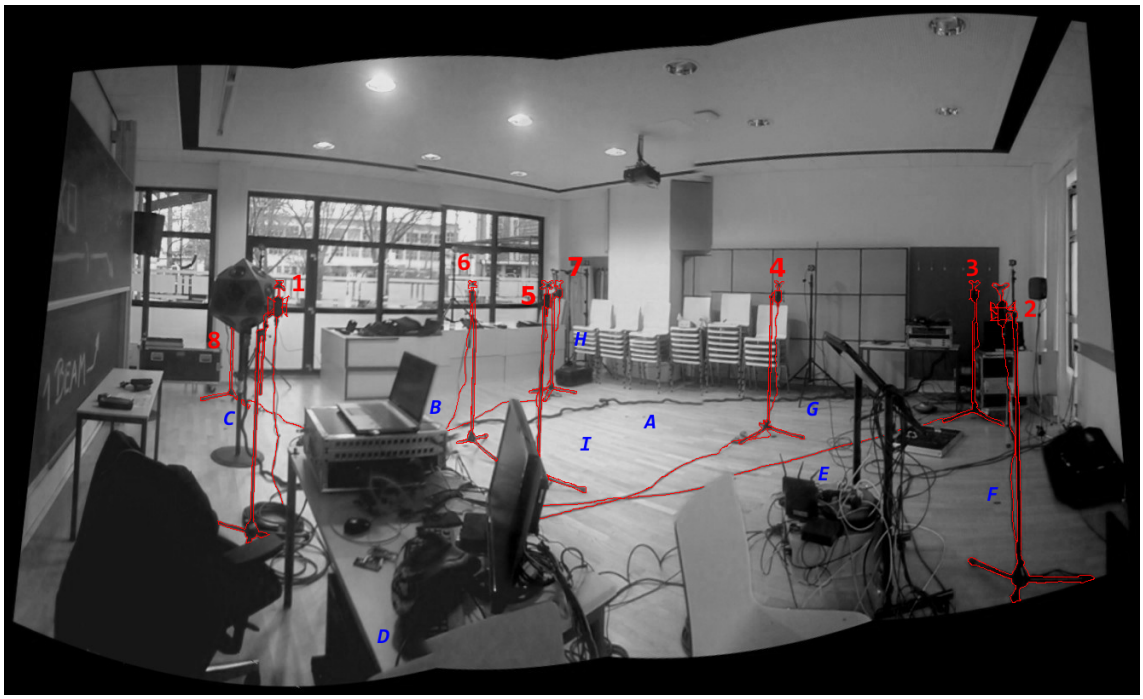


Figure 57: Setup for the second measurement, the positions of the `Oktava` microphones **1-8** are highlighted in red, the source positions where the claps are performed with **A-I** (projected onto the floor).

References

[CBH06] Jingdong Chen, Jacob Benesty, and Yiteng Huang. Time delay estimation in room acoustic environments: an overview. *EURASIP Journal on applied signal processing*, 2006:170–170, 2006.

[CDM12] Marco Crocco, Alessio Del Bue, and Vittorio Murino. A bilinear approach to the position self-calibration of multiple sensors. *IEEE Transactions on Signal Processing*, 60(2):660–673, 2012.

[CPSFC+14] Maximo Cobos, Juan J Perez-Solano, Santiago Felici-Castell, Jaume Segura, and Juan M Navarro. Cumulative-sum-based localization of sound events in low-cost wireless acoustic sensor networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(12):1792–1802, 2014.

[DDJ16] Mingjing Du, Shifei Ding, and Hongjie Jia. Study on density peaks clustering based on k-nearest neighbors and principal component analysis. *Knowledge-Based Systems*, 99:135–145, 2016.

[DDV16] Ivan Dokmani, Laurent Daudet, and Martin Vetterli. From acoustic room reconstruction to slam. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 6345–6349. IEEE, 2016.

[DV15] Ivan Dokmanic and Martin Vetterli. OMP with Unknown Filters for Multipath Channel Estimation. In *SPARS*, number EPFL-CONF-210571, 2015.

[DW86] B Dahanayake and K Wong. Proper orthogonal projection-multiple signal classification (pop-music). In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'86.*, volume 11, pages 2491–2494. IEEE, 1986.

[Fil13] Jason Filos. *Inferring Room Geometries*. PhD thesis, Imperial College London, 2013.

[Fit10] Derry Fitzgerald. Harmonic/percussive separation using median filtering. 2010.

[GB08] Michael Grant and Stephen Boyd. Graph implementations for non-smooth convex programs. In V. Blondel, S. Boyd, and H. Kimura, editors, *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, page 95–110. Springer-Verlag Limited, 2008. `http://stanford.edu/~boyd/graph_dcp.html`.

[GB14] Michael Grant and Stephen Boyd. CVX: Matlab software for disciplined convex programming, version 2.1. `http://cvxr.com/cvx`, March 2014.

[GKH13] Nikolay D Gaubitch, W Bastiaan Kleijn, and Richard Heusdens. Auto-localization in ad-hoc microphone arrays. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, page 106–110. IEEE, 2013.

[Hac15] Philipp Hack. Multiple source localization with distributed tetrahedral microphone arrays. Master's thesis, IEM, University of Music and Performing Arts Graz, 2015.

[HEK11] Mabande Haohai Sun, Kowalczyk E., and Kellermann K. Joint DOA and TDOA estimation for 3D localization of reflective surfaces using eigenbeam MVDR and spherical microphone arrays. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, page 113–116. IEEE, 2011.

[IK88] John Illingworth and Josef Kittler. A survey of the hough transform. *Computer vision, graphics, and image processing*, 44(1):87–116, 1988.

[KB14] Ian J. Kelly and Francis M. Boland. Detecting Arrivals in Room Impulse Responses with Dynamic Time Warping. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 22(7):1139–1147, July 2014.

[KC76] Charles Knapp and Glifford Carter. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(4):320–327, 1976.

[KSS13] Sarthak Khanal, Harvey F Silverman, and Rahul R Shakya. A free-source method (frsm) for calibrating a large-aperture microphone array. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(8):1632–1639, 2013.

[Kun96] Debasis Kundu. Modified music algorithm for estimating doa of signals. *Signal processing*, 48(1):85–90, 1996.

[KV96] Hamid Krim and Mats Viberg. Two decades of array signal processing research: the parametric approach. *Signal Processing Magazine, IEEE*, 13(4):67–94, 1996.

[LDW91] John J Leonard and Hugh F Durrant-Whyte. Mobile robot localization by tracking geometric beacons. *IEEE Transactions on robotics and Automation*, 7(3):376–382, 1991.

[LDWC92] John J Leonard, Hugh F Durrant-Whyte, and Ingemar J Cox. Dynamic map building for an autonomous mobile robot. *The International Journal of Robotics Research*, 11(4):286–298, 1992.

[LYMH11] Xuan Li, Shefeng Yan, Xiaochuan Ma, and Chaohuan Hou. Spherical harmonics music versus conventional music. *Applied Acoustics*, 72(9):646–652, 2011.

[Mes08] Xavier Mestre. On the asymptotic behavior of the sample estimates of eigenvalues and eigenvectors of covariance matrices. *IEEE Transactions on Signal Processing*, 56(11):5353–5368, 2008.

[ML99] John C Mosher and Richard M Leahy. Source localization using recursively applied and projected (rap) music. *IEEE Transactions on signal processing*, 47(2):332–340, 1999.

[ML08] Xavier Mestre and Miguel Ángel Lagunas. Modified subspace algorithms for doa estimation with large arrays. *IEEE Transactions on Signal Processing*, 56(2):598–614, 2008.

[NOS08a] F Nesta, M Omologo, and P Svaizer. Multiple tdoa estimation by using a state coherence transform for solving the permutation problem in frequency-domain bss. In *2008 IEEE Workshop on Machine Learning for Signal Processing*, page 43–48. IEEE, 2008.

[NOS08b] F Nesta, M Omologo, and P Svaizer. A novel robust solution to the permutation problem based on a joint multiple tdoa estimation. In *Proc. IWAENC*, 2008.

[Okt] `http://www.oktava-shop.com`, last access 14.11.2016.

[OS94] Maurizio Omologo and Piergiorgio Svaizer. Acoustic event localization using a crosspower-spectrum phase based technique. In *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*, volume 2, page II–273. IEEE, 1994.

[PDMP15] Archontis Politis, Symeon Delikaris-Manias, and Ville Pulkki. Direction-of-arrival and diffuseness estimation above spatial aliasing for symmetrical directional microphone arrays. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, page 6–10. IEEE, 2015.

[PDMPM15] D. Pavlidi, S. Delikaris-Manias, V. Pulkki, and A. Mouchtaris. 3D localization of multiple sound sources with intensity vector estimates in single source zones. In *Signal Processing Conference (EUSIPCO), 2015 23rd European*, page 1556–1560. IEEE, 2015.

[PN08] Marc Pollefeys and David Nister. Direct computation of sound and microphone locations from time-difference-of-arrival data. In *ICASSP*, page 2445–2448, 2008.

[RK89] R. Roy and T. Kailath. ESPRIT-estimation of signal parameters via rotational invariance techniques. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 37(7):984–995, 1989.

[SC78] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 26(1):43–49, 1978.

[Sch86] R.O. Schmidt. Multiple emitter location and signal parameter estimation. *Antennas and Propagation, IEEE Transactions on*, 34(3):276–280, 1986.

[SJHU+11] Joerg Schmalenstroeer, Florian Jacob, Reinhold Haeb-Umbach, Marius Hennecke, and Gernot A Fink. Unsupervised geometry calibration of acoustic sensor networks using source correspondences. In *Interspeech 2011*, 2011.

[SRZ10] N Shabtai, Boaz Rafaely, and Yaniv Zigel. Room volume classification from reverberant speech. In *Proc. of intl. Workshop on Acoustics Signal Enhancement, Tel Aviv, Israel*, 2010.

[SSC90] Randall Smith, Matthew Self, and Peter Cheeseman. Estimating uncertain spatial relationships in robotics. In *Autonomous robot vehicles*, page 167–193. Springer, 1990.

[STMK11] Haohai Sun, Heinz Teutsch, Edwin Mabande, and Walter Kellermann. Robust localization of multiple sources in reverberant environments

using eb-esprit with spherical microphone arrays. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 117–120. IEEE, 2011.

[SZR10] Noam R Shabtai, Yaniv Zigel, and Boaz Rafaely. Room volume classification from room impulse response using statistical pattern recognition and feature selection. *The Journal of the Acoustical Society of America*, 128(3):1155–1162, 2010.

[SZR13] Noam R Shabtai, Yaniv Zigel, and Boaz Rafaely. Towards room-volume classification from reverberant speech using room-volume feature extraction and room-acoustics parameters. *Acta Acustica united with Acustica*, 99(4):658–669, 2013.

[SZZ14] Christian Schörkhuber, Markus Zaunschirm, and IO-hannes Zmölnig. Wilma-wireless largescale microphone array. In *Linux Audio Conference*, volume 2014, 2014.

[TH13] O. Thiergart and E.A.P. Habets. An informed LCMV filter based on multiple instantaneous direction-of-arrival estimates. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, page 659–663. IEEE, 2013.

[TK10] Sakari Tervo and Teemu Korhonen. Estimation of reflective surfaces from continuous signals. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, page 153–156. IEEE, 2010.

[TKL11] Sakari Tervo, Teemu Korhonen, and Tapio Lokki. Estimation of reflections from impulse responses. *Building Acoustics*, 18(1-2):159–173, 2011.

[TP15] Sakari Tervo and Archontis Politis. Direction of arrival estimation of reflections from room impulse responses using a spherical microphone array. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 23(10):1539–1551, 2015.

[VTA+10] S Daniele Valente, Marco Tagliasacchi, Fabio Antonacci, Paolo Bestagini, Augusto Sarti, and Stefano Tubaro. Geometric calibration of distributed microphone arrays from acoustic source correspondences. In *Multimedia Signal Processing (MMSP), 2010 IEEE International Workshop on*, page 13–18. IEEE, 2010.

[Wei08] Stefan Weinzierl. *Handbuch der Audiotechnik*. Springer Science & Business Media, 2008.

[WIL] `http://wilma.kug.ac.at`, last access 14.11.2016.

[WVHK06] Earl G Williams, Nicolas Valdivia, Peter C Herdic, and Jacob Klos. Volumetric acoustic vector intensity imager. *The Journal of the Acoustical Society of America*, 120(4):1887–1897, 2006.

[YC08] Koh Yonghong Zeng and Ying-Chang Liang C.L. Maximum Eigenvalue Detection: Theory and Application. In *Communications, 2008. ICC '08. IEEE International Conference on*, page 4160–4164. IEEE, 2008.

[YGB13] Ehsan Yazdian, Saeed Gazor, and Mohammad Hasan Bastani. Limiting spectral distribution of the sample covariance matrix of the windowed array data. *EURASIP Journal on Advances in Signal Processing*, 2013(1):1–15, 2013.

[ZKS93] Michael D Zoltowski, Gregory M Kautz, and Seth D Silverstein. Beamspace root-music. *IEEE Transactions on Signal Processing*, 41(1):344, 1993.

Statutory Declaration

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

Graz, _____ _____

Date Signature

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Graz, am _____ _____

Datum Unterschrift