# Multiple Source Localization with Distributed Tetrahedral Microphone Arrays

Master Thesis by

**Philipp Hack**

Graz 2015

Host Institution:
Institute of Electronic Music and Acoustics
University of Music and Performing Arts Graz

Graz University of Technology

Assessor: Prof. Robert Höldrich
Supervisors: Dr. Franz Zotter
DI Christian Schörkhuber
DI Markus Zaunschirm

Institute of Electronic Music and Acoustics

## 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

# Acknowledgements

First of, i would like to thank my supervisors Christian Schörkhuber, Markus Zaunschirm and Dr. Franz Zotter for advising we with their profound knowledge, for their encouragement and for always having an open ear for me.
Further I am grateful to WKO Steiermark for providing me with a scholarship.

I want to dedicate this work to my parents and my girlfriend. Thank you for believing in me all this time, for your support, patience and motivation.

# Abstract

Microphone arrays have many different application areas and allow to analyze complex sound fields. This work focuses on acoustic source localization. In contrast to a central microphone array, which samples the sound field only locally and hence is limited to *Direction of Arrival* (DOA) estimation, distributed microphone arrays enable the estimation of absolute source positions in space. In this work, multiple distributed tetrahedral microphone arrays are used, where each tetrahedral array is able to estimate the DOA in 3 dimensions based on an *intensity vector* approach. By means of triangulation, the DOA estimates of the individual arrays can be integrated to a resulting source position estimate. Two different localization algorithms are developed, which are robust to sample clock synchronization mismatches among the arrays (important for wireless systems) and which provide the estimation of absolute source positions in 3D for single and multiple source scenarios. To evaluate the two algorithms real-world measurements are conducted, illuminating the effect of the number of used tetrahedral arrays on the localization performance. For optimal accuracy the DOA estimation performance of the Oktava 4D-Ambient tetrahedral arrays is improved by measurements and a DOA error-correction dictionary.

# Kurzfassung

Mikrofonarrays ermöglichen es, komplexe Schallfelder zu analysieren und besitzen vielfältige Anwendungsbereiche. Diese Arbeit beschäftigt sich mit dem Bereich der Schallquellenlokalisation. Während ein zentrales Mikrofonarray das Schallfeld nur lokal abtastet und sich damit beschränkt auf eine Schätzung der *Direction of Arrival* (DOA), besitzen verteilte Mikrofonarrays die Möglichkeit, absolute Schallquellpositionen zu schätzen. In dieser Arbeit werden mehrere räumlich verteilte Tetraedermikrofonarrays verwendet, welche über einen Intensitätsvektoransatz unabhängig voneinander 3D-DOA Schätzungen ermöglichen. Die geschätzten DOAs aller Tetraederarrays können z.B. über Triangulation zu einer resultierenden Schätzung der Quellpositionen einer oder mehrerer Schallquellen kombiniert werden. Es wurden 2 verschiedene Lokalisationsalgorithmen entwickelt, deren Lokalisation der absoluten Position einer oder mehrerer gleichzeitig aktiver Schallquellen im dreidimensionalen Raum robust gegenüber Sychronisationsfehlern zwischen den Array-Einheiten arbeitet (wichtig für drahtlose Systeme). Die beiden Lokalisationsalgorithmen werden anhand von Messungen in Abhängigkeit zur Anzahl verwendeter Tetraederarrays untersucht. Für eine optimale Genauigkeit in der DOA Schätzung der verwendeten Oktava 4D-Ambient Tetraederarrays verbessern Messungen und eine Datenbank zur DOA-Schätzungskorrektur die Ergebnisse.

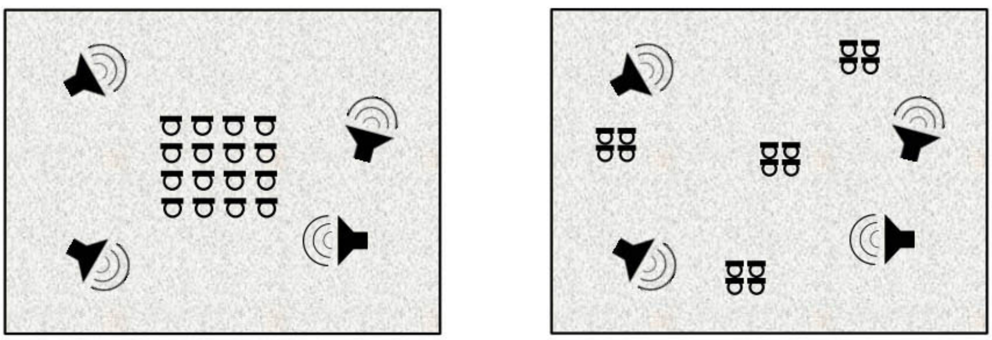
# Contents

# Chapter 1

# Introduction

Acoustic source localization (ASL) is the process of estimating the location of one or multiple sound sources in space by processing measurements of the underlying sound field [Fre10].

The most common ASL system we know are the human ears, which capture the soundfield at two discrete positions. The brain evaluates *Time Differences of Arrival* (TDOA) and *intensity differences* between the two ears as well as the spectral content of the captured sound and processes the information to an estimated source location. In engineering the measurements are typically conducted with an arrangement of two or more microphones, which is commonly referred to as a *microphone array*. By processing of the recorded microphone signals a source direction, also referred to as *Direction of Arrival* (DOA), and depending on the array configuration also absolute source locations can be estimated.



(a) Compact microphone array (CMA)　(b) Distributed microphone array (DMA)

Figure 1.1: Two different microphone array configurations.

For the purpose of DOA estimation typically *Compact Microphone Arrays* (CMAs) as illustrated in Fig. 1.1a are used. CMAs have been a topic of great interest for several decades and numerous DOA estimation strategies have been proposed. A typical application of CMAs is audio/video conferencing, where the goal is to steer a video camera or an acoustic beam former to the direction of the currently active speaker, which is estimated by the CMA [ZFDZ97, WC07]. Another application of ASL and a CMA is teleconferencing as in [APL07]. Here a monophonic audio stream of the recorded speaker signal with the estimated source direction of the speaker as side information is transmitted to reconstruct the source direction at the listeners side with a multiple loudspeaker setup. CMAs don't have access to distance information and they have difficulties estimating the absolute position of an acoustic source positioned in the array's far-field, as CMAs sample the sound field only locally, which can yield huge distances between the CMA and the source and thus poor SNR conditions.

In this work an alternative array configuration is used, the so called *Distributed Microphone Array* (DMA) shown in Fig. 1.1b, which allows us to estimate absolute source positions in 2- or 3-dimensional space as it covers a wider space of potential source locations compared to a CMA. It consists of several usually rather small microphone arrays that are spatially distributed in the volume of interest. Common microphone array localization algorithms can be applied to each individual CMA, allowing each CMA to work independently and in total, the CMAs cover multiple perspectives on a sound source. In a fusion center the information of all arrays is gathered and the arrays that provide the most promising information on the source location can be selected for ASL. In [TDdS+14] distributed arrays are used for noise pollution mapping in urban environments and in [BVRB10] for ambient assisted living, where the ASL estimations are used for beamforming, spatial representation of sounds by a multichannel playback system and controlling video systems or lights.

In contrast to what is dealt with in this work, the most popular ASL strategy is based on TDOA estimation. Here estimating the DOA or the absolute position of an acoustic source consists of two steps. For both cases first the TDOAs between the microphones (visible in the cross correlation between the microphone signals) are estimated. Popular algorithms for TDOA estimation are for example GCC-PHAT, which uses a frequency weighted version of the *Generalized Cross Correlation* (GCC) [KC76, HB04, BCH08] and AED, which stands for *Adaptive Eigenvalue Decomposition* and follows a channel identification approach [HBE99, HB04, BCH08]. As a second step the actual DOA estimation or source position estimation is done based on the estimated TDOAs, which yields a set of non-linear equations with no ideal closed form solution. Since no ideal closed form solution is available we have to search for a DOA or a point in space that fits to the measured TDOAs in an optimal way, i.e. define an error criterion and minimize it [BS97].

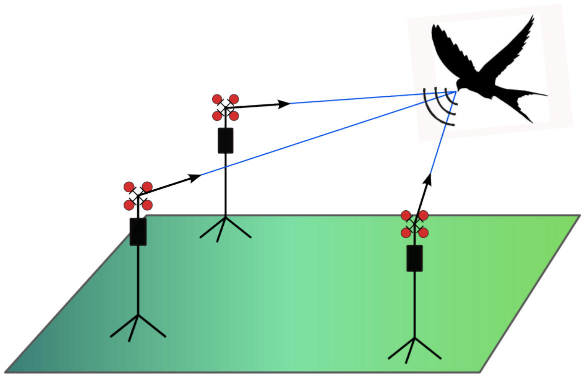


Figure 1.2: Simplified ASL example with 3 distributed tetrahedral microphone arrays [cli15].

The DMA used in this work consists of multiple distributed tetrahedral microphone arrays as illustrated in Fig. 1.2. The implemented ASL strategy differs from a TDOA based approach and is motivated in the next section. It is divided into two sub-tasks, of which the first task is DOA estimation. A big advantage of a tetrahedral microphone arrays is, that it allows to perform a 3-dimensional DOA estimation based on an *intensity vector* approach, which is part of the *Directional Audio Coding* (DirAC) algorithm presented in [PF06] and is further introduced in Ch. 2.1. DOA estimation is performed independently for each distributed tetrahedral microphone array in the frequency domain by applying a *Short Time Fourier Transform* (STFT). The second task: The individual arrays are not able to estimate the distance to the acoustic source, however in this work two different approaches are presented that estimate a source position in 3-dimensional space by combining the DOA estimations of all tetrahedral arrays. One approach (*linear intersection*) uses each DOA estimate as the direction vector of a bearing line that originates from the arrays position as shown in Fig. 1.2, i.e. a set of bearing lines is obtained. Ideally the estimated source position will then be the point in space where all bearing lines intersect. This concept is also known as *triangulation* and is commonly used in geodesy.

## 1.1 Motivation

In 2014 the *Wireless Large-Scale Microphone Array* (WiLMA) presented in [SZZ14] was developed by the *Institute of Electronic Music and Acoustics* (IEM). The WiLMA array was designed to analyze acoustic scenes in 3 dimensions under realistic conditions, including noise and reverberation. This also includes source parameter estimation like ASL, which was still an open topic and the initial motivation point for this work. The goal was to develop an ASL algorithm for multiple distributed spherical microphone arrays tailored to the WiLMA system.

The WiLMA system is a *wireless sensor network*, which consists of multiple distributed *Sensor Modules* (SM) that are each connected to 4 channel microphone arrays. The front-end of each SM provides 4 channels with high-end preamps and A/D-converters ($f_s = 48$ kHz, 24 Bit). After conversion the signals of each SM are multiplexed and transmitted over a wireless or wired link to a central unit, where the signal processing takes place. In addition every SM is equipped with a local processing unit that can do simple pre-analysis of the signals and thus reduce the computational load on the central unit and also reduce the amount of data that has to be transmitted. On the microphone side each SM is connected to an Oktava 4D-Ambient tetrahedral 4 channel microphone array [SZZ14]. Fig. 1.3 shows a potential acoustic scene analysis scenario with 3 distributed WiLMA arrays.

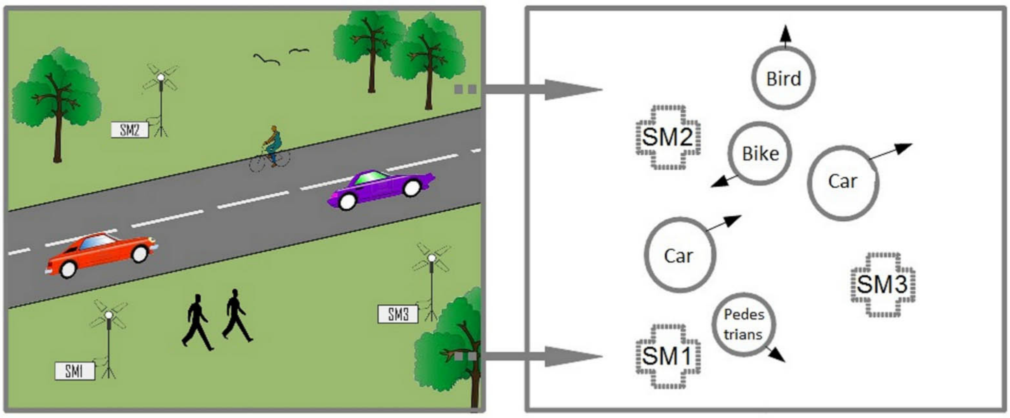



Figure 1.3: Acoustic scene analysis with a WiLMA system consisting of 3 spatially distributed sensor modules (SM) [SZZ14].

The system can be used e.g. for source tracking and source identification shown in the right picture of Fig. 1.3 or source extraction. All these applications demand a robust source position estimate. An issue of a wireless transmission system is that time synchronization cannot always be guaranteed between all SM. Therefore an important criterion was to design an ASL algorithm that is robust to slight time delays between the sensor nodes, i.e. time critical TDOA based localization strategies could not be used. Since the Oktava 4D-Ambient microphones were not tested yet, a second goal of this work was to evaluate the characteristics of the Oktava microphone arrays and their qualification for DOA estimation.

## 1.2 Thesis Outline

Chapter 2, *DOA Estimation with a Single Tetrahedral Microphone Array*, introduces the geometry of a tetrahedral microphone array and describes how DOA estimation is performed and DOA histograms are generated with such an array. By means of a mathematical model the properties of a tetrahedral array are investigated.

Chapter 3, *Source Localization in 3 Dimensions with Multiple Microphone Arrays*, presents two developed ASL algorithms that combine the DOA estimations and histograms of multiple individual tetrahedral arrays and integrate them to a resulting source position estimation in 3-dimensional space.

Chapter 4, *Measuring the Tetrahedral Oktava Microphone Array*, evaluates characteristics and the DOA estimation performance of a single tetrahedral microphone array based on real-world measurements.

Chapter 5, *Comparison of Model and Measurement*, properties and DOA estimation performance of mathematical model and measurements are compared for a single tetrahedral microphone array.

Chapter 6, *DOA Correction*, the measurements from Ch.5 are used to correct the DOA estimation performance of the tetrahedral microphone array.

Chapter 7, *Experimental Evaluation*, the two localization algorithms are evaluated by means of real-world measurements.

Chapter 8, *Conclusion and Outlook*

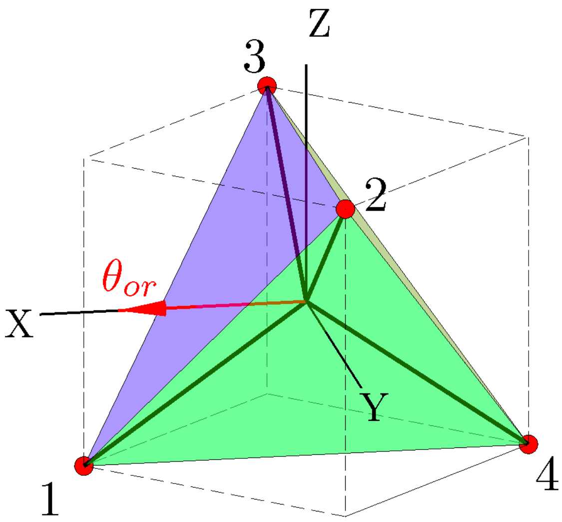
# Chapter 2

# DOA Estimation with a Single Tetrahedral Microphone Array

The localization of absolute source positions is divided into 2 sub-tasks, where first a DOA estimation is performed locally on each tetrahedral array and then the DOA estimates of multiple tetrahedral arrays are combined to obtain an estimate of the absolute source positions in 3-dimensional space. In this chapter DOA estimation with a single tetrahedral microphone array based on the *intensity vector* approach is presented and the Oktava 4D-Ambient array is introduced as the tetrahedral microphone array of choice. An ideal model of the microphone array is derived, which will be compared in Ch. 4.5.3 and Ch. 5 to the acoustic measurements of the Oktava 4D-Ambient array that are described in Ch. 4.

Fig. 2.1a shows the Oktava 4D-Ambient microphone array and the arrangement of the 4 microphones in a tetrahedral design, where the microphones are placed in the diagonally opposite corners of a cubes front and backside. Connecting the 4 microphone positions through lines leads to 4 triangular faces, which meet at each vertex and form the regular tetrahedrons body [BH09, Bat09].

(a) Oktava 4D-Ambient




(b) Tetrahedral microphone positions

Figure 2.1: Tetrahedral microphone configuration and definition of the array orientation vector $\boldsymbol{\theta}_{or}$.

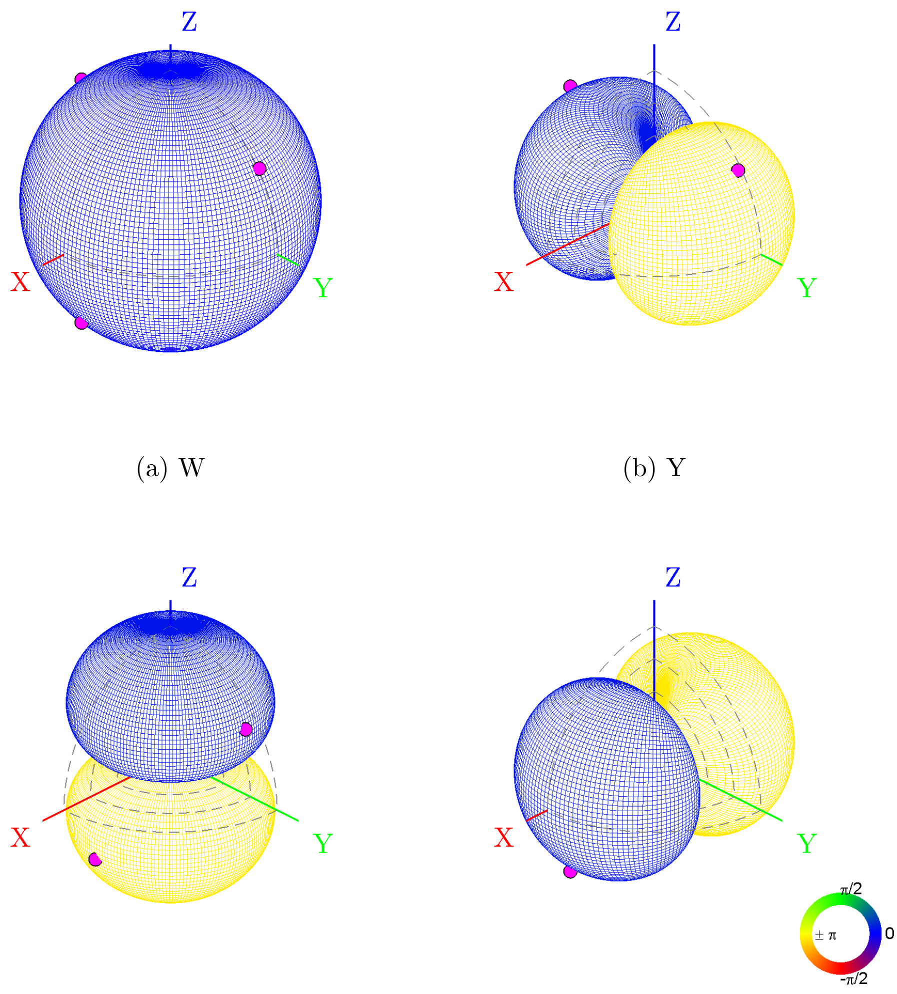
Expressing the microphone positions in spherical coordinates $\mathbf{r}_j = (R, \varphi_j, \vartheta_j)$ with a radius $R = 3.5$ cm of the Oktava 4D-Ambient array yields the following values for angles $\varphi_j$ and $\vartheta_j$

| j | $\varphi_j$ in ° | $\vartheta_j$ in ° |
|---|---|---|
| 1 | 0 | $90 + 35.264$ |
| 2 | 90 | $90 - 35.264$ |
| 3 | 270 | $90 - 35.264$ |
| 4 | 180 | $90 + 35.264$ |

Table 2.1: Microphone positions.

Compared to traditional convention as in [Far79, BH09] the microphone positions used in this work (see Tab. 2.1) are rotated by $\Delta\varphi = +45°$ around the z-axis, which yields a different A to B-Format conversion matrix in Ch. 2.2. This is done for practical reasons, since microphone 1 could be used as an indicator to align the array's orientation vector $\boldsymbol{\theta}_{or}$ illustrated in Fig. 2.1b with the x-axis during the measurements conducted in Ch. 4. The orientation vector describes the orientation of the microphone array and is defined as the projection of position $\mathbf{r}_1$ onto the array's horizontal plane.

Using vector notation the 4 microphone signals can be written as $\boldsymbol{s} = (s_1,\ s_2,\ s_3,\ s_4)^{\mathrm{T}}$ (*A-Format*). Multiplication with an encoder matrix $\mathbf{Y}_{N_A}^{-1}$ defined in Eq. 2.8 transforms the sampled A-Format signals into the spherical harmonics domain and yields $\boldsymbol{\chi}_{N_A} = \mathbf{Y}_{N_A}^{-1}\,\boldsymbol{s}$, where $\boldsymbol{\chi}_{N_A} = (\mathrm{W},\ \mathrm{Y},\ \mathrm{Z},\ \mathrm{X})^{\mathrm{T}}$ (*B-Format*) denotes the spherical wave spectrum of zeroth and first order which will be needed for DOA estimation. $N_A$ indicates the highest spherical harmonics order that can be extracted by the microphone array, which is $N_A = 1$ for the tetrahedral array. The A- to B-Format conversion is described in detail in Ch. 2.2. Fig. 2.2 illustrates the ideal directivity patterns (no spatial aliasing, ideal first order microphones) of the 4 B-Format signals W,Y,Z and X.




(a) W

(b) Y

(c) Z

(d) X

Figure 2.2: Directivity patterns of B-Format components W, Y, Z and X and microphone positions (•).

## 2.1 Intensity Vector Approach

Using the B-Format signals, the DOA estimation can be accomplished by a simple intensity vector approach [MP05, APL07]. The instantaneous sound intensity vector $\boldsymbol{I} = p\mathbf{v}$ requires the knowledge of sound pressure $p$ and the particle velocity vector $\mathbf{v} = (\mathrm{v}_x, \mathrm{v}_y, \mathrm{v}_z)^\mathrm{T}$ at a given point. The B-Format signals yield an estimate of both at the center point of the microphone array, where W estimates $p$ and $\mathrm{X}, \mathrm{Y}, \mathrm{Z}$ estimate $(\mathrm{v}_x, \mathrm{v}_y, \mathrm{v}_z)\ \varrho_0 c$, with the density of air $\varrho_0$ and the velocity of sound $c$. As the intensity vector indicates the magnitude and the direction of the transport of acoustical energy of a sound wave, its inverse direction can be directly used as a DOA estimate [JHN10].

To enable DOA estimation of multiple (concurrent) sources based on the intensity vector approach we make use of the temporal and spectral disjointness property exhibited by most sound mixtures, which is explained later. Since the spectral disjointness property is observable in frequency domain, time-frequency analysis is applied to the B-Format signals by first windowing the B-Format signals with a Hann window of 93 $ms$ length and then performing a *Discrete Fourier Transform* (DFT) of the windowed frames. This is done every time instant and yields the complex B-Format vector $\boldsymbol{\chi}_N(n, \mathrm{k}) = (\mathrm{W}[n, \mathrm{k}],\ \mathrm{Y}[n, \mathrm{k}],\ \mathrm{Z}[n, \mathrm{k}],\ \mathrm{X}[n, \mathrm{k}])^\mathrm{T}$, with time index $n$ and frequency index k, $\mathrm{k} = 1, \ldots, \mathrm{K}$. The frequency corresponding to frequency index k is denoted by $f_\mathrm{k}$. For the sake of better readability the time index $n$ is omitted from now on. The instantaneous intensity vector can then be calculated individually for each frequency index k with[1] [MP05, APL07]

$$\boldsymbol{I}_\mathrm{k} = \frac{1}{\varrho_0 c}\, \Re \left\{ \mathrm{W}^*[\mathrm{k}] \begin{pmatrix} \mathrm{X}[\mathrm{k}] \\ \mathrm{Y}[\mathrm{k}] \\ \mathrm{Z}[\mathrm{k}] \end{pmatrix} \right\}. \tag{2.1}$$

The instantaneous DOA estimate at frequency index k is expressed by the unit vector [SHZ+14]

$$\hat{\boldsymbol{\theta}}_\mathrm{k} = -\frac{\boldsymbol{I}_\mathrm{k}}{\|\boldsymbol{I}_\mathrm{k}\|}, \tag{2.2}$$

where $\|\cdot\|$ is a vector's $\ell_2$ norm. For the localization of multiple concurrent sources the property of spectral disjointness is exploited. Spectral disjointness means that the spectral contents of most acoustic sources will be distributed differently over the

[1] To calculate the intensity vector in Eq. 2.1, the dipole patterns of the first-order spherical harmonics components $\mathrm{X}, \mathrm{Y}, \mathrm{Z}$ have to point in the opposite direction to the coordinate axes [JHN10]. The signs of the dipoles illustrated in Fig. 2.2 are adjusted accordingly.

entire spectrum such that the spectral content at each frequency $f_{\mathrm{k}}$ is dominated by a single source or noise. E.g. the speech signals of two different speakers will have different fundamental frequencies and thus also different harmonics. If a 2-dimensional histogram is generated over $\hat{\varphi} \in [0, 2\pi)$ and $\hat{\vartheta} \in [0, \pi)$ from the estimated DOAs in spherical coordinates $\hat{\boldsymbol{\theta}}_{\mathrm{k}} \equiv (\hat{\varphi}_{\mathrm{k}}, \hat{\vartheta}_{\mathrm{k}})$ at all K frequencies of a DFT frame, then even concurrent active sources will generate distinctive peaks approximately at the real source directions $\varphi$, $\vartheta$ if they are spectrally disjoint.[2] The generation of the 2-dimensional histograms will now be explained in detail.

### 2.1.1 DOA Histograms

Assuming spectral disjointness, resulting DOA estimates for each source can be obtained through the computation of a 2-dimensional DOA histogram $\boldsymbol{h}$ (see Fig. 2.4). The data used to create the histogram are the K DOA estimates from one or multiple DFT frames. To generate $\boldsymbol{h}$, a grid of $Q$ directions $\{\boldsymbol{\uptheta}_q\}_Q$ is defined on the unit sphere and stored in a codebook. In a search process the direction $\boldsymbol{\uptheta}_{\check{q}_{\mathrm{k}}}$ from the codebook is determined that is closest to an estimated DOA $\hat{\boldsymbol{\theta}}_{\mathrm{k}}$. The index $\check{q}_{\mathrm{k}}$ corresponding to the closest codebook direction is indicated through a binary detector (Kronecker delta) $\delta_{q,\check{q}_{\mathrm{k}}}$ (Eq. 2.4). A histogram entry $h[q]$ now corresponds to the number of DOA estimates $\hat{\boldsymbol{\theta}}_{\mathrm{k}}$ that are mapped to the codebook direction with index $q$. But rather than just counting the DOA estimates that correspond to $h[q]$, each estimate is weighted by a signal strength factor $g_{\mathrm{k}}$

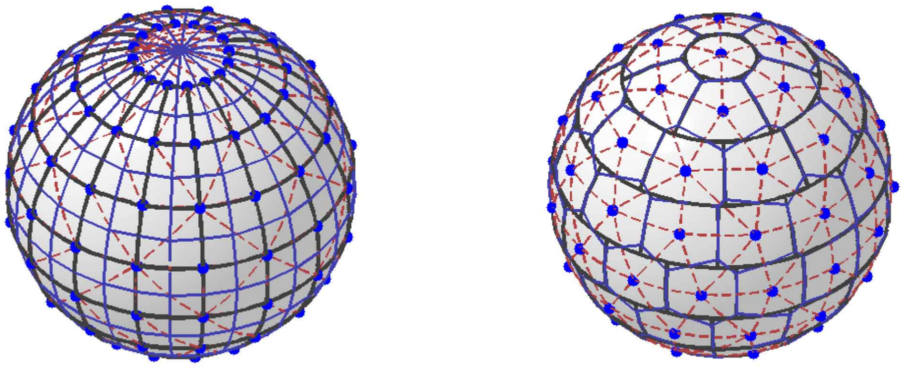
$$
\begin{aligned}
h[q] &= \sum_{\mathrm{k}} g_{\mathrm{k}}\ \delta_{q,\check{q}_{\mathrm{k}}}, \\
\check{q}_{\mathrm{k}} &= \arg\min_{q} \|\boldsymbol{\uptheta}_q - \hat{\boldsymbol{\theta}}_{\mathrm{k}}\|, \\
g_{\mathrm{k}} &= |\mathrm{W}[\mathrm{k}]|^{\alpha}, \quad \text{with } \alpha \geq 0
\end{aligned}
\tag{2.3}
$$

with the binary detector

$$
\delta_{q,\check{q}_{\mathrm{k}}} = \begin{cases} 1 & \text{if} \quad q = \check{q}_{\mathrm{k}}, \\ 0 & \text{otherwise.} \end{cases}
\tag{2.4}
$$

[2]If e.g. two sources are not spectral disjointed at frequency index k then a direction between the real source directions will be estimated as the DOA.

The weighting with the signal strength factor $g_{\mathrm{k}}$ follows the idea that frequency indices k with high signal amplitude generate more reliable DOA estimates than those with small amplitude. The absolute value of the omni-directional B-Format component W[k] is exponentiated with $\alpha$. The size of $\alpha$ (recommended value $\alpha = 1$) determines how important large signal amplitude is rated in the histogram generation.



(a) Gaussian sampling, $Q = 128$ (b) equal area sampling, $Q = 75$

Figure 2.3: Two different sampling schemes, with sampled positions $\boldsymbol{\theta}_q$ (blue dots) and Voronoï cells (areas confined by blue lines) on the unit circle. [Zot09]

There are various ways to sample the sphere at $Q$ directions [Pom08, Raf05, Zot09]. In this work the *Gaussian sampling* scheme is used, which samples the azimuth angle $\varphi$ at $2(N+1)$ and the zenith angle $\vartheta$ at $N+1$ equi-angular samples . Thus the angle stepsize between neighboring samples is equal in azimuth and zenith direction and for instance is chosen to $\Delta\varphi = \Delta\vartheta = 1°$, which results in 360 azimuth samples and 180 zenith samples and amounts to a total number of $Q = 64800$ codebook directions $\boldsymbol{\theta}_q$. As an estimated direction $\hat{\boldsymbol{\theta}}_{\mathrm{k}}$ is assigned to the nearest direction sample $\boldsymbol{\theta}_{\tilde{q}_{\mathrm{k}}}$, each sampled direction is surrounded by an area on the sphere which it represents. The area is called a *Voronoï cell*. As shown in Fig. 2.3a for the Gaussian sampling scheme the Voronoï cells are not of equal size, respectively the cells are largest at the equator and become smaller towards the poles. Because of this, the entries of the 2-dimensional histograms will be slightly biased towards the equator, since larger Voronoï cells capture more estimated directions $\hat{\boldsymbol{\theta}}_{\mathrm{k}}$. An alternative sampling scheme distributes the sampling directions $\boldsymbol{\theta}_q$ in a such a way that the sphere is partitioned into equal areas, i.e. all Voronoï cells are of the same size (see Fig. 2.3b). The biasing effect immanent for the Gaussian sampling scheme is thereby avoided, however as the biasing error for the Gaussian sampling scheme is assumed to be small, the Gaussian sampling scheme is used [SHZ+14].

Fig. 2.4 shows the 2-dimensional histograms $\boldsymbol{h}$ computed from Eq. 2.3 with codebook directions $\boldsymbol{\theta}_q$ based on the Gaussian sampling scheme for a scenario of two sources. In Fig. 2.4a,b,c the sources are two female speakers, whereas in Fig. 2.4d the two sources radiate white noise. Fig. 2.4a shows the histogram $\boldsymbol{h}$ if the value of the weighting exponent is chosen to $\alpha = 0$. The two peaks are well separated which proves, that the assumption of spectral disjointness holds. Since the two peaks exhibit several spurious side peaks, which can make peak picking difficult, the histograms are further filtered[3] with a 2-dimensional Gaussian low pass filter, which is defined as

$$H(\varphi, \vartheta) = \frac{1}{2\pi\sigma^2}\, e^{-\frac{\varphi^2 \,+\, \vartheta^2}{2\sigma^2}}, \tag{2.5}$$

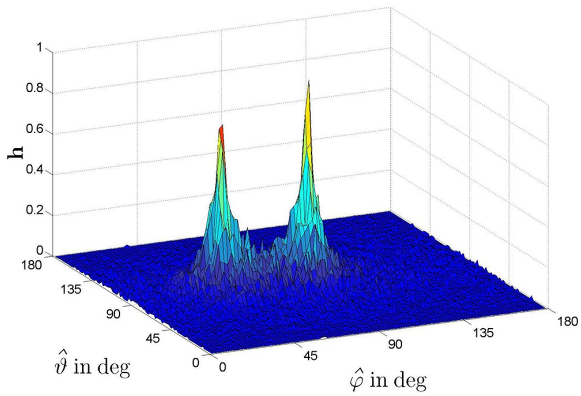
where the standard deviation is chosen to $\sigma = 0.76$.
The smoothed histogram is illustrated in Fig. 2.4b. For Fig. 2.4a,b no signal strength weighting is applied, i.e. $\alpha = 0$. This is different for Fig. 2.4c where the weighting coefficient is set to $\alpha = 1$. As a consequence an increase in the sharpness of the peaks can be observed, which is beneficial for estimation accuracy and for the differentiation of neighboring peaks. In Fig. 2.4d the two sources radiate white noise and therefore the assumption of approximate spectral disjointness does not hold here. Consequently the DOA estimation results in an erroneous single peak which is located in between the DOAs of the two sources.

## 2.2 A- to B-Format Conversion

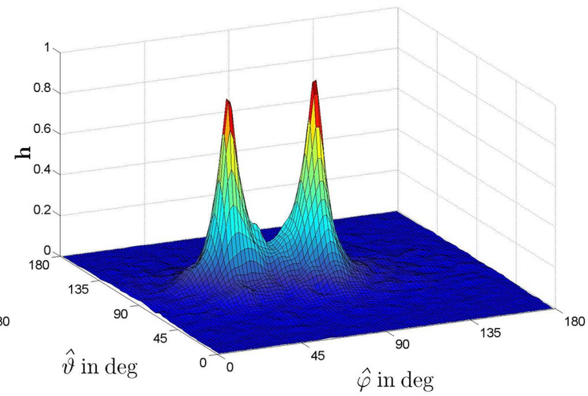
The signals $\boldsymbol{s} = (s_1[n], s_2[n], s_3[n], s_4[n])^{\mathrm{T}}$ that are directly recorded through the 4 microphones constitute the so called A-Format, where $n$ is the time index. We use the notation $\boldsymbol{s}$ instead of $\boldsymbol{p}$, since we use cardioid microphones and therefore do not measure the pressure $p$, but a mixture of pressure and velocity (Eq. 2.9).[4]

Through the 4 used microphones, the sound field on the sphere is sampled at 4 discrete positions. With Eq. A.17 the sampled soundfield $\boldsymbol{s}$ on a sphere can be decomposed into a set of spherical harmonics, which yields a system of linear equations that is solvable for $J \geq (N+1)^2$, where $J$ is the number of microphones and $N$ the spherical harmonics order. Generally for $(N+1)^2 = J$ the system of linear equations is fully

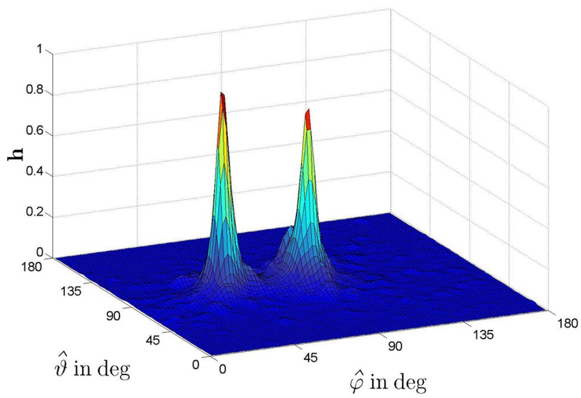
[3]Filtering corresponds to *spherical convolution*.
[4]The corresponding spherical wave spectrum is denoted by $\chi_n^m$. (see Eq. 2.6)



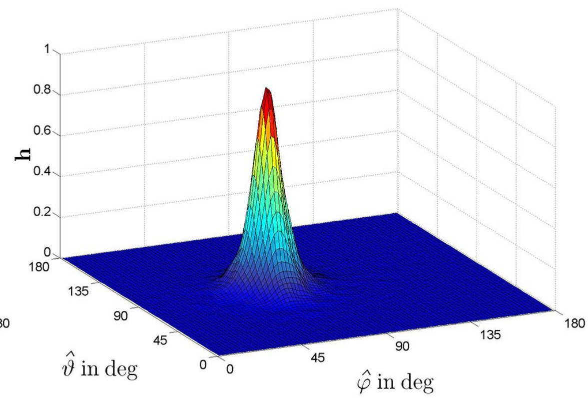

(a) speech, $\alpha = 0$ and not filtered




(b) speech, $\alpha = 0$ and Gauss filtered




(c) speech, $\alpha = 1$ and Gauss filtered




(d) white noise, $\alpha = 1$ and Gauss filtered

Figure 2.4: Histograms $\boldsymbol{h}$ generated with Gaussian sampling for two active sources at positions $(1.5\,\text{m},\ 45°,\ 90°)$ & $(1.5\,\text{m},\ 90°,\ 90°)$ with and without Gauss filtering and weighting $\alpha$. The histograms are normalized to the magnitude of their max. peak and shown over the range $\hat{\vartheta} \in [0\ 180°)$ and $\hat{\varphi} \in [0\ 180°)$ (second half $\hat{\varphi} \in [180\ 360°)$ of the histograms is not shown for the sake of better readability). For (a,b,c) the sources are two female speakers and for (d) white noise.

determined, therefore to obtain a solution for Eq. A.17 using the pseudo inverse is not necessary as $\mathbf{Y}_N$ is square and can be inverted directly (provided that $\mathbf{Y}_N$ is well conditioned). The tetrahedron however belongs to the group of *platonic solids* that if used as a sampling scheme have the convenient property of orthonormal spherical harmonics, i.e. $\mathbf{Y}_N \mathbf{Y}_N^{\mathrm{T}} = a\mathbf{I}$. Thus $\mathbf{Y}_N^{-1}$ can be easily computed as $\mathbf{Y}_N^{-1} = \frac{1}{a}\mathbf{Y}_N^{\mathrm{T}}$ without any matrix inversion. Written in the frequency domain, Eq. 2.6 decomposes the pressure distribution on the sphere of radius $R$ into spherical harmonics, where $k = {}^{\omega}\!/_{c}$ denotes the wavenumber (not to be confused with frequency index k) with c being the speed of sound. In case of the considered tetrahedral microphone array the extractable spherical harmonics order is $N = N_A = 1$, i.e. decomposed spherical harmonics are of zeroth and first order.

$$\boldsymbol{s}(k) = \mathbf{Y}_{N_A} \boldsymbol{\chi}_{N_A}(kR) \quad \rightarrow \quad \boldsymbol{\chi}_{N_A}(kR) = \mathbf{Y}_{N_A}^{-1} \boldsymbol{s}(k) = \frac{1}{a} \mathbf{Y}_{N_A}^{\mathrm{T}} \boldsymbol{s}(k), \tag{2.6}$$

where $\boldsymbol{\chi}_{N_A}$ is the spherical wave spectrum and $a = \frac{1}{\pi}$. With the spherical harmonics $Y_n^m(\boldsymbol{\theta}_j)$ evaluated at the microphone directions in sph. coordinates $\boldsymbol{\theta}_j \equiv (\varphi_j, \vartheta_j)$, $j = 1, \ldots, 4$ (Tab. 2.1) and written in vector notation as in Eq. A.18 yields

$$\mathbf{y}_1(\varphi_j, \vartheta_j) = \left( \tfrac{1}{\sqrt{4\pi}},\ \sqrt{\tfrac{3}{4\pi}} \sin\varphi_j \sin\vartheta_j,\ \sqrt{\tfrac{3}{4\pi}} \cos\vartheta_j,\ \sqrt{\tfrac{3}{4\pi}} \cos\varphi_j \sin\vartheta_j \right)^{\mathrm{T}}, \tag{2.7}$$

and

$$\mathbf{Y}_1^{-1} = \pi \mathbf{Y}_1^{\mathrm{T}} \approx \begin{pmatrix} 0.886 & 0.886 & 0.886 & 0.886 \\ 0 & -1.253 & 1.253 & 0 \\ -0.886 & 0.886 & 0.886 & -0.886 \\ 1.253 & 0 & 0 & -1.253 \end{pmatrix}, \tag{2.8}$$

with $\mathbf{Y}_1^{\mathrm{T}} = (\mathbf{y}_1(\boldsymbol{\theta}_1), \mathbf{y}_1(\boldsymbol{\theta}_2), \mathbf{y}_1(\boldsymbol{\theta}_3), \mathbf{y}_1(\boldsymbol{\theta}_4))$. The spherical wave spectrum $\boldsymbol{\chi}_1$ contains the four B-Format signals $\boldsymbol{\chi}_1 = \left[\chi_0^0,\ \chi_1^{-1},\ \chi_1^0,\ \chi_1^1\right]^{\mathrm{T}} = [\mathrm{W},\ \mathrm{Y},\ \mathrm{Z},\ \mathrm{X}]^{\mathrm{T}}$, where the coefficient $\chi_0^0$ of the zeroth order spherical harmonic corresponds to an omni-directional pressure signal W and the first order coefficients $\chi_1^{-1},\ \chi_1^0,\ \chi_1^1$ correspond to dipole signals Y, Z, X along the orthogonal y, z and x axes.

Let us take a closer look at $\mathbf{Y}_1^{-1}$ in Eq. 2.8. From Eq. 2.6 one can see, that e.g. B-Format component X is generated by the inner product of the fourth row of $\mathbf{Y}_1^{-1}$ with $\boldsymbol{s}$. Since the row has only two nonzero values, only $s_1$ and $s_4$ are used to generate X. This is reasonable as $\varphi_1$ and $\varphi_4$ (see Tab. 2.1) point in $\pm x$ direction, while $\varphi_2$ and $\varphi_3$ point in $\pm y$ direction and thus do not contribute to the X dipole signal.

### 2.2.1 Analytical Model

In this chapter an analytic model of the tetrahedral microphone array will be derived to observe how the spherical wave spectrum $\boldsymbol{\chi}_{N_A}$ can be analytically estimated from the sampled soundfield signals $\mathbf{s}$ on a sphere with radius $R = 3.5$ cm, where the soundfield is described by the wave spectrum $\boldsymbol{b}_{N_C}$ of higher order ($N_C > N_A$). The analytic model is based on ideal first order directional microphones described by Eq. 2.9, implying ideal conditions with no directivity pattern or gain mismatches between the microphones, no microphone positioning errors and the on axis directions of the microphones equal the ideal tetrahedral positions given in Tab. 2.1 for all frequencies. On the basis of the analytic model, the cardioid microphone configuration will be compared to two alternative configurations. Further the effects of spatial aliasing will be shown.

#### 2.2.1.1 Radial Functions and Filters

The output of a single first order directional microphone can be mathematically described as the summation of an omni-directional and a bidirectional microphone signal. Written in frequency domain we get [BH09, Bat09]

$$s(\mathbf{r},k) = \beta\ p(\mathbf{r},k)\ -\ (1-\beta)\varrho_0 c \mathrm{v}_r(\mathbf{r},k), \tag{2.9}$$

where $\mathbf{r} = R\,\boldsymbol{\theta}$ is a point on the sphere with radius $R$, $p$ is the pressure, $\mathrm{v}_r$ the radial particle velocity, $k$ the wavenumber, constant $\varrho_0$ denotes the specific density of air and $\beta$, $0 \leq \beta \leq 1$ is a parameter that controls the type of first order directivity pattern. For $\beta = 0$ a purely omni-directional and for $\beta = 1$ a purely bidirectional pattern is obtained [Fre10].[5] Insertion of the Euler equation for the radius coordinate

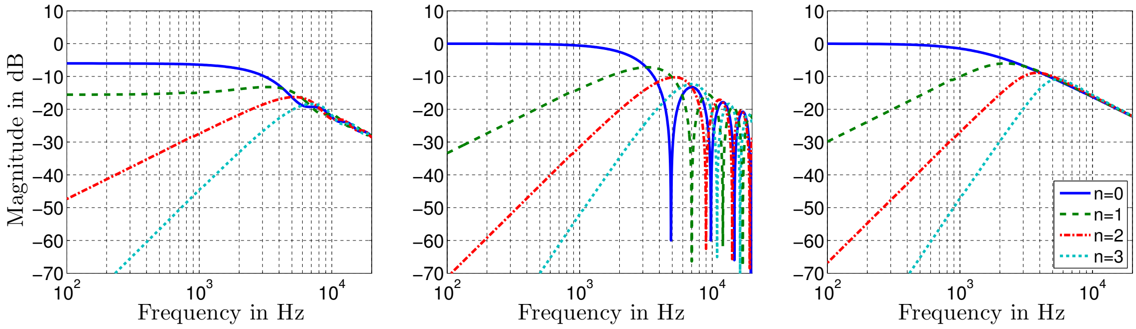
$$i\varrho_0 c\ k\mathrm{v}_r(\mathbf{r},k) = \frac{\partial\ p(\mathbf{r},k)}{\partial r}, \tag{2.10}$$

into Eq. 2.9 and using Eq. A.17, which decomposes the distribution of $s(\mathbf{r},k)$ on the sphere of radius $R$ into the spherical harmonics domain, yields the coefficients of the spherical wave spectrum as

$$\begin{aligned}
\chi_n^m(kR) &= \int_0^{2\pi}\int_0^{\pi} s(\mathbf{r},k) Y_n^m(\varphi,\vartheta)\sin(\vartheta) d\vartheta\ d\varphi && (2.11)\\
&= \int_0^{2\pi}\int_0^{\pi} \left(\beta\ p(\mathbf{r},k) - i(1-\beta)\ \frac{1}{k}\ \frac{\partial\ p(\mathbf{r},k)}{\partial r}\right) Y_n^m(\varphi,\vartheta)\sin(\vartheta) d\vartheta\ d\varphi \\
&= \beta\ b_n^m j_n(kR) - i(1-\beta)\ \frac{1}{k}\ \frac{\partial j_n(kR)}{\partial r} b_n^m \\
&= \underbrace{\left(\beta\ j_n(kR) - i(1-\beta)\ j_n'(kR)\right)}_{\rho_n(kR)} b_n^m, && (2.12)
\end{aligned}$$

[5] Directivity patterns and *Directivity Index* (DI) for different $\beta$ values: [Fre10]
omni $\rightarrow \beta = 1$, ($DI = 0$ dB) | figure-eight $\rightarrow \beta = 0$, ($DI = 4.77$ dB) | cardioid $\rightarrow \beta = 0.5$, ($DI = 4.77$ dB) | supercardioid $\rightarrow \beta = 0.37$, ($DI = 5.72$ dB) | hypercardioid $\rightarrow \beta = 0.25$, ($DI = 6.02$ dB).

where $\rho_n(kR) = \beta\ j_n(kR) - i(1-\beta)\ j_n'(kR)$ are the radial functions for the cardioid open-sphere scenario and $j_n$ denote the spherical Bessel functions. The wave spectrum $b_n^m$, which characterizes the wave field at one frequency, can be obtained by multiplying the spherical wave spectrum $\chi_n^m(kR)$ with the radial filters $V_n(kR) = {}^1/_{\rho_n(kR)}$. The magnitude response of the radial functions $\rho_n(kR)$ is illustrated in Fig. 2.5a for a radius of $R = 3.5$ cm and different orders $n = 0, \ldots, 3$.

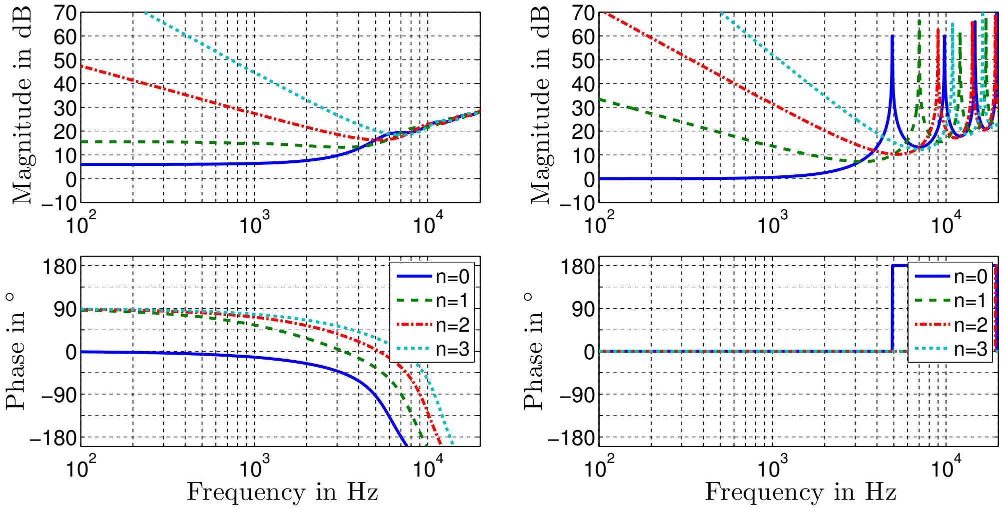



(a) Open-sphere cardioid (b) Open-sphere omni (c) Rigid sphere omni

Figure 2.5: Magnitude of radial functions $\rho_n(kR)$ for orders $n = 0, 1, 2, 3$ and a radius of $R = 3.5$ cm.

Replacing $\rho_n(kR)$ of Eq. 2.11 with the two other radial functions given in Tab. A.1 yields the spherical wave spectrum of two alternative types of spherical array configurations, namely the open-sphere configuration using omni-directional microphones [APL07, BB07] and the rigid-sphere configuration, where the surface of a rigid sphere is sampled with omni-directional microphones [LD05, BB07]. For comparison Fig. 2.5b and Fig. 2.5c show the magnitude responses of the radial functions for the open-sphere and closed-sphere scenario.

For all three scenarios in Fig. 2.5 the spherical harmonic of zeroth order has a flat amplitude response up to approximately 2 kHz. For the two scenarios with omni-directional microphones spherical harmonics of order $n \geq 1$ show a high-pass characteristic with an increase in magnitude of $20 \cdot n$ dB/Dec., whereas for the scenario with cardioid microphones the orders $n \geq 1$ start out with higher magnitudes at low frequencies and increase with $20 \cdot (n-1)$ dB/Dec. [Zau12]. The less agressive filtering is provided by the cardioid transducers as they acoustically derive signals related to velocity, i.e. the first-order spatial gradient integrated over time, whereas $\frac{1}{\rho_n(kR)}$ needs to incorporate this step when using omnidirectional transducers.




(a) open-sphere cardioid

(b) open-sphere omni

Figure 2.6: Frequency response of radial filters $V_n(kR)$ for orders $n = 0, 1, 2, 3$ and a radius of $R = 3.5$ cm.

The open-sphere configuration with omni-directional transducers has another disadvantage compared to the cardioid scenario: The factor $\frac{1}{\rho_n(kR)}$ causes singularities because of the zero crossings of the spherical Bessel functions $j_n(kR)$ (see Fig. 2.6b). By its moderate $\frac{1}{\rho_n(kR)}$ factor, the favorized open-sphere cardioid transducer array is most robust for first-order decomposition, i.e. the robustness to errors in microphone positioning and sensor self noise is increased. Nevertheless, cardioid arrays are more challenging in hardware than in signal processing: Cardioid microphones are usually less precise in terms of frequency response matching, self-noise, and in how well they achieve the prescribed directivity [BB07].

#### 2.2.1.2 Spatial Aliasing

Depending on the sampling scheme a soundfield can be decomposed into spherical harmonics of order $N_A \leq N_{max} = \sqrt{J} - 1$, where $J$ are the number of sampling points. We can only extract the spherical harmonics of a sound field up to order $N_A$. As already mentioned for the tetrahedral array holds $N_A = N_{max} = 1$. If a continuous function $s(\mathbf{r})$ is sampled on the sphere at the discrete positions of the sampling scheme (into the vector $\boldsymbol{s}$) and is not strictly band-limited to orders below $N_A$, the linear system in Eq. 2.6 is under-determined and there exist infinitely many solutions (aliases) $\boldsymbol{\chi}_{N_C}$ that lead to the same values of $\boldsymbol{s}$. This ambiguity is typically referred to as spatial aliasing [Pom08]. On the other hand, the finite-order $N_A$ yields a limited spatial resolution of the microphone array, which is recognized in the shape of the array's beampattern, with its wide lobes [Raf05].

Exciting a soundfield with finite-order $N_C$, Eq. 2.11 can be written in matrix vector notation with Eq. 2.6. The sampled soundfield is obtained as

$$\boldsymbol{s} = \mathbf{Y}_{N_C}\mathrm{diag}\{\boldsymbol{\rho}_{N_C}(kR)\}\boldsymbol{b}_{N_C}. \tag{2.13}$$

Applying Eq. 2.6 to Eq. 2.13 describes the decomposition of the sampled function $s(\mathbf{r}, k)$ on a sphere of radius $R$ into spherical harmonics up to order $N_A$ on the microphone side. The function $s(\mathbf{r}, k)$ on the sphere is excited by an underlying wave spectrum $\boldsymbol{b}_{N_C}$ of order $N_C$ (excitation side), where $N_C \geq N_A$:

$$\underbrace{\boldsymbol{\chi}_{N_A} = \mathbf{Y}^{+}_{N_A}}_{\text{microphone side}}\overbrace{\mathbf{Y}_{N_C}\mathrm{diag}\{\boldsymbol{\rho}_{N_C}(kR)\}\boldsymbol{b}_{N_C}}^{\text{exitation side}}, \tag{2.14}$$

If $\mathbf{s}$ is band-limited with order $N_A$, i.e. $N_C = N_A$ the analytic model is free of spatial aliasing and its decomposition works perfectly, i.e. $\boldsymbol{\chi}_{N_A} = \mathrm{diag}_{N_C}\{\boldsymbol{\rho}_{N_C}\}\boldsymbol{b}_{N_C}$. For $N_C > N_A$ spatial aliasing occurs and the spherical wave spectrum $\boldsymbol{\chi}_{N_A}$ is corrupted by the higher order spherical harmonics $N_A < n \leq N_C$ of $\boldsymbol{b}_{N_C}$ that are interpreted as lower-order spherical harmonics with $n \leq N_A$. Since spherical harmonic magnitudes of sound on the sphere decay for $n > kr$, cf. Fig. 2.5, an upper frequency $f_u$ can be defined from $n = kr$ as

$$f_u = \frac{n\ c}{2\pi R}, \tag{2.15}$$

above which we can expect spatial aliasing. For the Oktava 4D-Ambient microphone array with $n = N_A = 1$ this frequency lies at $f_u \approx 1560$ Hz.

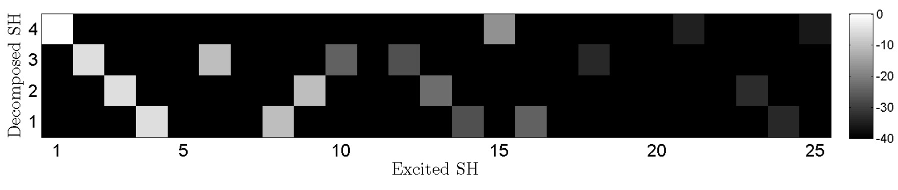
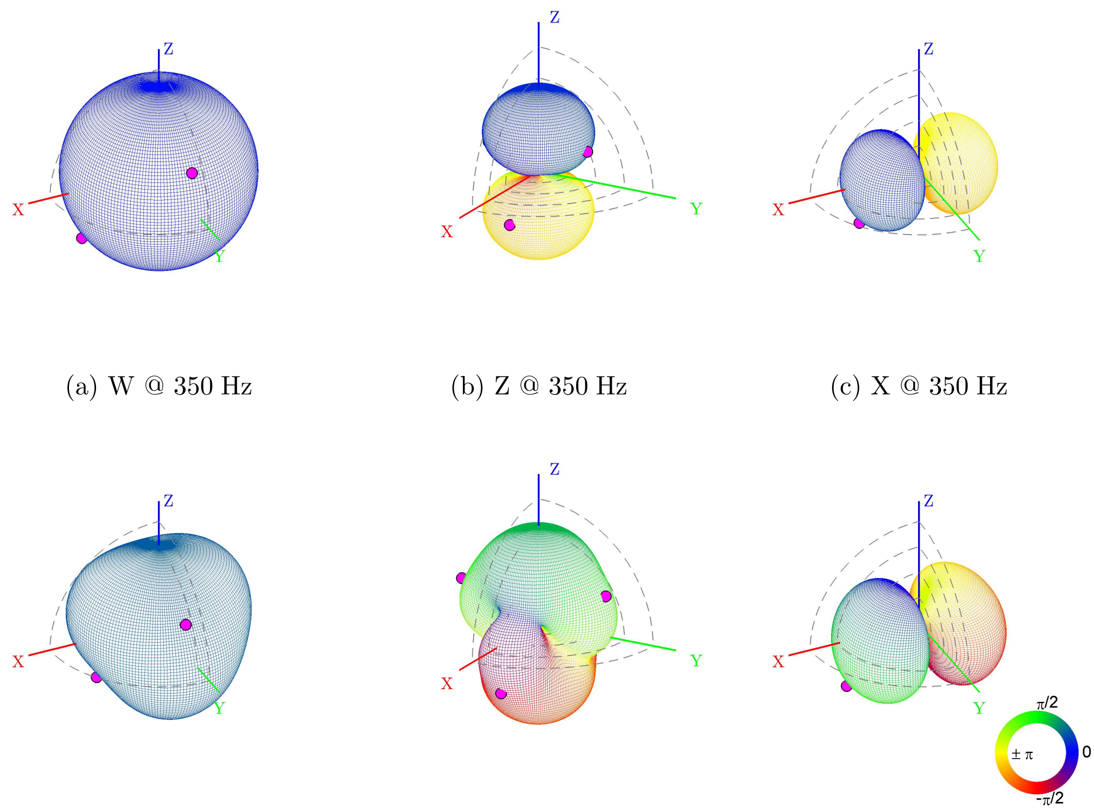



Figure 2.7: First 25 columns of system matrix $\dot{\mathbf{M}}_{N_A,N_C}$ at a frequency of 2000 Hz

Spatial aliasing can be regarded in greater detail by inspecting the absolute values in the wave spectrum to B-format matrix $\dot{\mathbf{M}}_{N_A,N_C} = \mathbf{Y}^{+}_{N_A}\mathbf{Y}_{N_C}\mathrm{diag}_{N_C}\{\boldsymbol{\rho}_{N_C}(kR)\}$ from

Eq. 2.14, cf. Fig. 2.7, which should ideally equal the identity matrix zero-padded with columns on the right. The matrix is referred to as the system matrix of the analytic model. The columns correspond to the $(N_C + 1)^2$ spherical harmonic components on the excitation side and the rows correspond to the $(N_A + 1)^2$ decomposed components on the microphone side. Fig. 2.7 shows the first 25 columns of the matrix. The left-most 4 columns indicate the array-achievable subspace showing perfect mapping of the wave spectrum $\boldsymbol{b}_1$ to the B-format $\boldsymbol{\chi}_1$ without crosstalk or scaling. The other columns indicate spatial aliasing, to which, obviously second-order spherical harmonics deliver the greatest contribution at 2 kHz, in particular spherical harmonics 6,8 and 9. Spatial aliasing increases with frequency and is dominated by second-order components throughout the entire relevant frequency range.




(a) W @ 350 Hz (b) Z @ 350 Hz (c) X @ 350 Hz

(d) W @ 2000 Hz (e) Z @ 2000 Hz (f) X @ 2000 Hz

Figure 2.8: Directivity patterns of the analytic model based B-Format components Z and X of $\boldsymbol{\chi}_{N_A}$ at different frequencies. (• → microphone positions).

The $9^{th}$ spherical harmonic for example is interpreted as a Z component of the B-Format Signals $\boldsymbol{\chi}_{N_A}$. Its influence can be recognized in the directivity pattern Fig. 2.8e by comparison to the directivity pattern of the $9^{th}$ spherical harnomic illustrated in

Fig. A.4. By contrast, for a frequency of 350 Hz Fig. 2.8b shows an almost ideal figure of eight pattern.

Since the shape of the second-order harmonics that map to the horizontal dipole patterns, see 6 and 8 in Fig. A.4 are aligned with the x and y directions, their influence on the B-format's X and Y signals are less apparent, but similarly strong. What is more, because directional sampling is the reason for spatial aliasing, it is natural that the particular influence of spatial aliasing depends on the rotation of the array.

# Chapter 3

# Source Localization in 3 Dimensions with Multiple Microphone Arrays

By combination of the DOA estimation data obtained from multiple tetrahedral microphone arrays it is possible to localize multiple acoustic sources in 3-dimensional space. In this chapter two different localization algorithms are proposed, the *acoustic map* algorithm which is presented in the following chapter and the *linear intersection* algorithm described in Ch. 3.2.

## 3.1 Acoustic Map Algorithm

The acoustic map algorithm is based on the DOA histograms $\boldsymbol{h}^m$, $m = 0, 1, \ldots, M$ of $M$ tetrahedral microphone arrays (see Ch. 2.1.1). The arrays are distributed in 3-dimensional space and their positions are indicated in a global cartesian coordinate system by $\mathbf{r}_m$. The volume of potential source locations is rastered into a grid of a discrete set of points $\mathbf{g}_p$, $p = 1, \ldots, P$ where $P$ is the total number of points. For each of these grid points $\mathbf{g}_p$ we calculate $M$ difference vectors

$$\mathbf{v}_{p,m} = \mathbf{g}_p - \mathbf{r}_m. \tag{3.1}$$

The difference vectors are normalized to unit vectors

$$\tilde{\mathbf{\Phi}}_{p,m} = \frac{\mathbf{v}_{p,m}}{\|\mathbf{v}_{p,m}\|}, \tag{3.2}$$

where $\|\cdot\|$ is a vector's $\ell_2$ norm. For each array we then define a local coordinate system according to Fig. 2.1 from Ch. 2, where the on-axis microphone orientations are given by Tab. 2.1. The unit vectors $\tilde{\mathbf{\Phi}}_{p,m}$ in Eq. 3.2, defined in the global coordinate

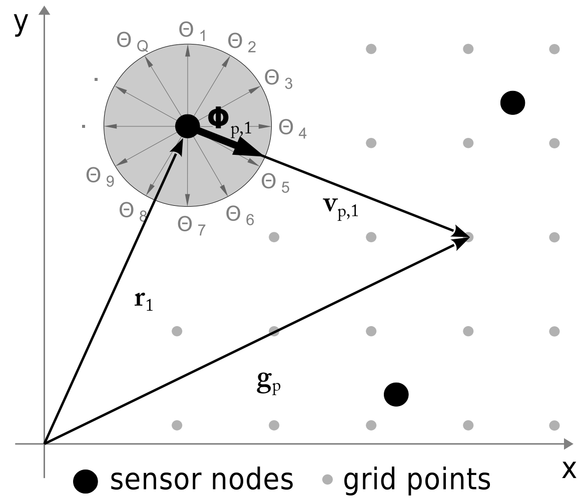



Figure 3.1: 2D-illustration of the difference vector $\mathbf{v}_{p,m}$ between a grid point $\mathbf{g}_p$ and array position $\mathbf{r}_m$. Further the mapping of $\mathbf{v}_{p,m}$ to the nearest code book direction $\boldsymbol{\theta}_{\check{q}_{p,1}}$ is illustrated, which in this case is $\boldsymbol{\theta}_5$. [SHZ+14]

system, are transformed into the local coordinate system of each tetrahedral array by applying an array-dependent rotation matrix $\mathbf{R}_m$:

$$\boldsymbol{\Phi}_{p,m} = \mathbf{R}_m \; \tilde{\boldsymbol{\Phi}}_{p,m}. \tag{3.3}$$

By doing that, we obtain the direction of a grid point from the $m^{th}$ microphone array's point of view. The rotation matrix considers rotations between the global and the local coordinate systems. If the local and the global coordinate system align, the rotation matrix becomes the unitary matrix $\mathbf{R}_m = \mathbf{I}$ and $\boldsymbol{\Phi}_{p,m} = \tilde{\boldsymbol{\Phi}}_{p,m}$.
From the histogram $\boldsymbol{h}^m$ the entry $h^m[\check{q}_{p,m}]$ is chosen that corresponds to the codebook position $\boldsymbol{\theta}_{\check{q}_{p,m}}$ which is closest to $\boldsymbol{\Phi}_{p,m}$ by

$$\check{q}_{p,m} = \arg\min_q \|\boldsymbol{\theta}_q - \boldsymbol{\Phi}_{p,m}\|, \tag{3.4}$$

where $\check{q}_{p,m}$ denotes the index corresponding to the closest codebook position. For each grid point $\mathbf{g}_p$ we can define a column vector $\mathbf{b}_p$ that contains the histogram entries $h^m[\check{q}_{p,m}]$ of all $M$ arrays, which yields

$$\mathbf{b}_p = \begin{pmatrix} h^1[\check{q}_{p,1}] \\ h^2[\check{q}_{p,2}] \\ \vdots \\ h^M[\check{q}_{p,M}] \end{pmatrix}. \tag{3.5}$$

That is, the histogram $\boldsymbol{h}^m$ of each array is projected into space and evaluated at the discrete grid points $\mathbf{g}_p$. According to Eq. 3.5, for each grid point we get $M$ histogram values. In a next step these $M$ values are combined by superposition, which leads to a first version of the spatial likelihood function or acoustic map

$$\mathcal{L}'(\mathbf{g}_p) = \mathbf{b}_p^{\mathrm{T}} \cdot \mathbf{1}_{M\times 1}, \tag{3.6}$$

where the scalar product with the unitary column vector $\mathbf{1}_{M\times 1}$ simply sums up the $M$ elements of $\mathbf{b}_p$. $\mathcal{L}'(\mathbf{g}_p)$ describes the likelihood of a source being present at grid point $\mathbf{g}_p$. Thus the elements of $\mathbf{b}_p$ indicate the contribution of each array's histogram to the total likelihood of a present source (see Fig. 3.3a).

### 3.1.1 Normalization and Weighting

In this chapter we will introduce the normalization of the histograms $\boldsymbol{h}^m$ and the *sparsity weighting* function, which supplies an individual weighting factor for each grid point $\mathbf{g}_p$. Both aim to reduce two undesired side effects which are caused by the underlying nature of the proposed localization strategy. The two effects are referred to as *array dominance* and *ghost peaks*. Further *distance degradation* is presented as a second subsequent weighting scheme.

**Array Dominance:** This effect occurs if for a grid point $\mathbf{g}_p$ an element $h^m[\check{q}_{p,m}]$ of $\mathbf{b}_p$ dominates the other elements in amplitude. The reason for this is a peak of high amplitude in the corresponding array's histogram $\boldsymbol{h}^m$ compared to the peaks in the histograms of the other arrays. This can be due to an active source positioned in close vicinity to an array. A close source will be present with a relatively high amplitude level in the array's output signals compared to the other sources. This dominance of the close source in signal level is also prominent in the frequency domain and thus most DOA estimates $\hat{\boldsymbol{\theta}}_k$ point in the direction of the close source and lead to a dominating peak in the corresponding histogram $\boldsymbol{h}^m$, where $m$ is the index of the array close to the source. Array dominance can also be caused by very consistent DOA estimates that lead to narrow peaks of high amplitude. If the direction corresponding to the dominating peak in $\boldsymbol{h}_m$ does not match the direction of the true source position it will change the resulting peak of that source in the acoustic map significantly due to its dominating character.

**Normalization:** To reduce the array dominance effect, two different procedures are applied. The first one is a normalization of the histograms $\boldsymbol{h}_m$. The purpose of the normalization procedure is to bring the peaks in all histograms $\boldsymbol{h}^m$ closer together in amplitude, i.e. to compress the peak levels so that high peak amplitudes are still emphasized, but only up to a certain degree in order to prevent the array dominance effect. This is realized by first identifying the maximum $h^m_{max}$ of each histogram $\boldsymbol{h}^m$, $m = 1, \ldots, M$ and then calculating the mean value

$$\bar{h}_{max} = \frac{1}{M} \sum_{m=1}^{M} h^m_{max}. \tag{3.7}$$

To force the peak amplitudes $h^m_{max}$ to be closer to $\bar{h}_{max}$, a simple compression function $\mathcal{T}(.)$ is used. The modified histograms are obtained by

$$\tilde{\boldsymbol{h}}^m = \underbrace{\frac{1}{\tilde{h}^m_{max}} \mathcal{T}\left(\tilde{h}^m_{max}\right)}_{\kappa^m} \boldsymbol{h}^m, \tag{3.8}$$

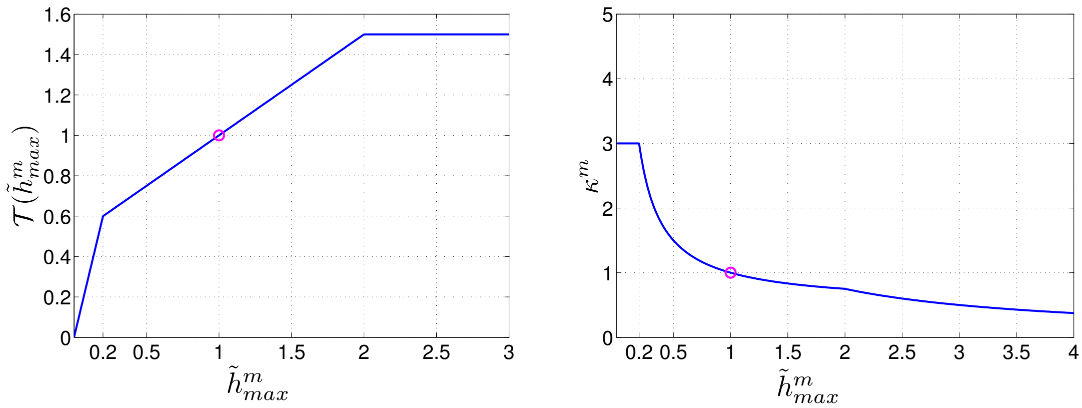
where $\tilde{h}^m_{max} = {h^m_{max}}/{\bar{h}_{max}}$ and $\kappa^m$ is the resulting correction factor corresponding to the $m^{th}$ microphone array. The compression function illustrated in Fig. 3.2a is defined as

$$\mathcal{T}(x) = \begin{cases} 3 \cdot x & \text{if} \quad x < 0.2, \\ 0.5 \cdot x + 0.5 & \text{if} \quad 0.2 \leq x \leq 2, \\ 1.5 & \text{if} \quad x > 2. \end{cases} \tag{3.9}$$

With Eq. 3.6 a second version of the acoustic map is obtained by

$$\mathcal{L}''(\mathbf{g}_p) = \tilde{\mathbf{b}}_p^{\mathrm{T}} \cdot \mathbf{1}_{M \times 1}, \tag{3.10}$$

where $\tilde{\mathbf{b}}_p$ corresponds to the $\mathbf{b}_p$ vector generated from the normalized histograms $\tilde{\boldsymbol{h}}^m$.




(a) Compression function $\mathcal{T}(.)$

(b) Correction factor $\kappa^m$

Figure 3.2: Compression function $\mathcal{T}(\tilde{h}^m_{max})$ and correction factor $\kappa^m$.

**Sparsity Weighting:** To further reduce the array dominance effect as a second procedure an individual weighting factor is applied to each grid point $\mathbf{g}_p$. At the position of a real source ideally all arrays make some contribution to the peak value in the acoustic map, i.e. the vector $\tilde{\mathbf{b}}_p$ contains multiple significant entries. For grid points $\mathbf{g}_p$ where the contribution of a single array dominates this is not the case. Utilizing this observation we apply a penalty function $\mathcal{P}(\tilde{\mathbf{b}}_p)$ to $\mathcal{L}''(.)$, which decreases with increasing sparsity of $\tilde{\mathbf{b}}_p$. Sparsity is lowest if the contributions of the arrays in $\tilde{\mathbf{b}}_p$ are distributed uniformly, e.g. for four arrays $\tilde{\mathbf{b}}_p = [1/4,\ 1/4,\ 1/4,\ 1/4]^T$ and highest if only one element is non zero, e.g. $\tilde{\mathbf{b}}_p = [0,\ 0,\ 1,\ 0]^T$. We define

$$\mathcal{P}(\tilde{\mathbf{b}}_p) = \frac{1}{\mathcal{G}(\tilde{\mathbf{b}}_p) + \epsilon}, \tag{3.11}$$

where $\mathcal{G} \in (0, 1)$ is the *Gini index* and $\epsilon$ is a parameter that controls the impact of $\mathcal{G}$ with $\epsilon > 0$ to avoid singularities. The Gini index is a measure of sparsity that is commonly used to assess the inequality in income or wealth and is defined as [HR08]

$$\mathcal{G}(\mathbf{c}) = 1 - 2\sum_{l=1}^{L} \frac{\tilde{c}_{(l)}}{||\tilde{\mathbf{c}}||_1} \left(\frac{L - l + 0.5}{L}\right), \tag{3.12}$$

where $\tilde{\mathbf{c}}$ is a sorted version of $\mathbf{c}$ such that

$$\tilde{c}_{(1)} \leq \tilde{c}_{(2)} \leq \cdots \leq \tilde{c}_{(L)}.$$

Applying the sparsity weights from Eq. 3.11 to Eq. 3.6 a third version of the acoustic map is obtained as

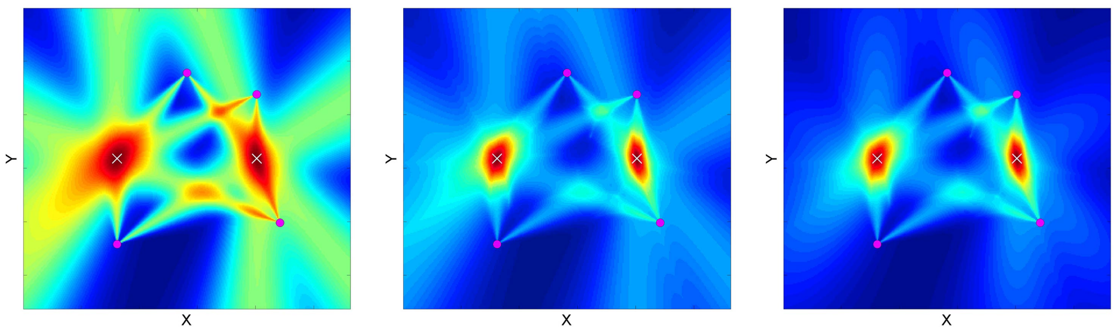
$$\mathcal{L}'''(\mathbf{g}_p) = \tilde{\mathbf{b}}_p^{\mathrm{T}} \cdot \mathbf{1}_{M\times 1}\mathcal{P}(\tilde{\mathbf{b}}_p). \tag{3.13}$$

Note that $\mathcal{P}(\tilde{\mathbf{b}}_p)$ is a scalar weighting factor, i.e. there is only one factor for each grid point $\mathbf{g}_p$. Note that sparsity weighting should only be used in moderation, since most acoustic sources exhibit focused radiation patterns for higher frequencies, which results in an increased sparsity of $\tilde{\mathbf{b}}_p$.

**Ghost Peaks:** The second undesired effect are *ghost peaks*. Ghost peaks occur when multiple sources are localized with multiple distributed microphone arrays. Fig. 3.5 shows the *localization lines* of each microphone array for a scenario of three microphone arrays and two active sources. The localization lines represent the peak directions of the DOA histograms $\boldsymbol{h}^m$, $m = 1, 2, 3$ projected into 2-dimensional space.

As illustrated in Fig. 3.5, ghost peaks are caused by superposition wherever two localization lines cross at positions where no real sources occur. Depending on the scenario these peaks can exhibit higher values in the acoustic map than the peaks corresponding to real sources. Fig. 3.5 shows that mostly two localization lines, i.e. two microphone arrays are involved wherever a ghost peak appears. Thus except for two significant elements, all other elements in the vector $\tilde{\mathbf{b}}_p$ will be close to zero, if a ghost peak is present at position $\mathbf{g}_p$. Therefore, the sparsity weighting from Eq. 3.11 will also reduce the impact of ghost peaks, as it applies less weight to grid points with high sparsity. The effect of ghost peaks is further reduced by the *acoustic map de-emphasis*, which is described in detail in Ch. 3.1.3.

An exemplary simulated acoustic map $\mathcal{L}'''(\mathbf{g}_p)$ is illustrated in Fig. 3.3b. The normalization and sparsity weighting applied in $\mathcal{L}'''(\mathbf{g}_p)$ lead to a huge improvement compared to $\mathcal{L}'(\mathbf{g}_p) = \mathbf{b}_p^{\mathrm{T}} \cdot \mathbf{1}_{M\times 1}$ in Fig. 3.3a. All regions in the acoustic map that only receive contributions from few arrays are weighted less, which yields a much more distinct acoustic map in Fig. 3.3b.




(a) $\mathcal{L}'(\mathbf{g}_p) = \mathbf{b}_p^{\mathrm{T}} \cdot \mathbf{1}_{M\times 1}$ (b) $\mathcal{L}'''(\mathbf{g}_p) = \tilde{\mathbf{b}}_p^{\mathrm{T}} \cdot \mathbf{1}_{M\times 1}\mathcal{P}(\tilde{\mathbf{b}}_p)$ (c) $\mathcal{L}(\mathbf{g}_p) = \tilde{\mathbf{b}}_p^{\mathrm{T}}\mathbf{d}_p \cdot \mathcal{P}(\tilde{\mathbf{b}}_p)$

Figure 3.3: Illustration of theoretical (simulated) 2D acoustic maps without and with different weightings, with 4 microphone arrays (•) and 2 sources (×).

**Distance Degradation Weighting:** In Fig. 3.3b the projections of each histograms into space still spread out until the boundaries of the observed volume are reached. Observing an arbitrary grid point $\mathbf{g}_p$ we want to weight the contributions in $\tilde{\mathbf{b}}_p$ from arrays that are close to $\mathbf{g}_p$ more and contributions from distant arrays less, since we expect more reliable DOA estimations from arrays which are close. One reason for this assumption is, that the cone of high probability of an array in Fig. 3.3b gets wider with increasing distance and therefore distant arrays generate unfocused peaks in the acoustic map. Weighting is achieved by applying a function $\mathcal{D}(\|\mathbf{v}_{p,m}\|) \in (0,1)$ that decreases with increasing distance between $\mathbf{g}_p$ and $\mathbf{r}_m$ as in [TDdS+14]

$$\mathcal{D}(\|\mathbf{v}_{p,m}\|) = e^{-\frac{\|\mathbf{v}_{p,m}\|^2}{\lambda}}, \tag{3.14}$$

where $\lambda$ denotes the *decay factor*. By writing the $M$ distance weights that belong to each grid point in a vector notation as

$$\mathbf{d}_p = \begin{pmatrix} \mathcal{D}(\|\mathbf{v}_{p,1}\|) \\ \mathcal{D}(\|\mathbf{v}_{p,2}\|) \\ \vdots \\ \mathcal{D}(\|\mathbf{v}_{p,M}\|) \end{pmatrix}, \tag{3.15}$$

the weighting of each element $h^m[\check{q}_{p,m}]$ of $\tilde{\mathbf{b}}_p$ with the corresponding weight $\mathcal{D}(\|\mathbf{v}_{p,m}\|)$ can be realized by an inner product, which yields the final version of the acoustic map

$$\mathcal{L}(\mathbf{g}_p) = \tilde{\mathbf{b}}_p^{\mathrm{T}} \mathbf{d}_p \cdot \mathcal{P}(\tilde{\mathbf{b}}_p). \tag{3.16}$$

The effect of the distance weighting is illustrated in Fig. 3.3c, where $\mathcal{L}(\mathbf{g}_p)$ is simulated in 2 dimensions. The projections of the histograms $\boldsymbol{h}^m$ into space now possess a limited radius of influence.

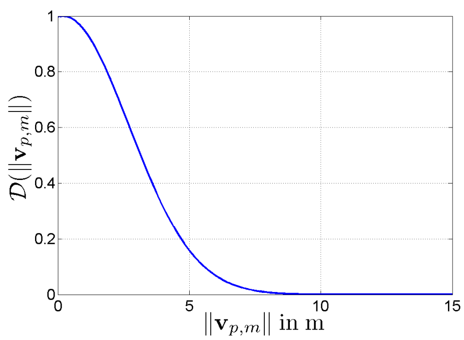



Figure 3.4: Distance degradation function $\mathcal{D}(\|\mathbf{v}_{p,m}\|)$ for $\lambda = 12.5$.

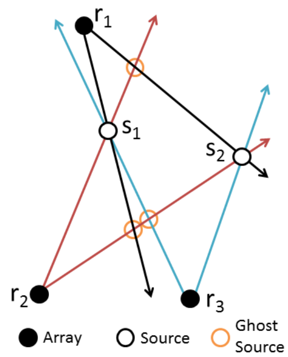



Figure 3.5: Appearance of ghost peaks.

The size of this radius is determined by the decay factor $\lambda$, that has to be chosen according to the distances between the microphone arrays. Fig. 3.4 shows the distance function itself for $\lambda = 12.5$. It should be noted here that it is important to apply the distance degradation after the sparsity weighting, otherwise the peaks of more distant sources will be reduced two times; first by the distance degradation and then also by the sparsity weighting since the distance degradation causes the elements corresponding to arrays with huge distance from the source to vanish and thus increases sparsity.

### 3.1.2 Finding the Source Positions

To localize active acoustic sources in 3 dimensions, the weighted 3-dimensional acoustic map $\mathcal{L}(\cdot)$ is searched for its global maximum over all grid points $\mathbf{g}_p$, $p = 1, \ldots, P$.

$$\hat{\mathbf{s}} = \arg \max_{\mathbf{g}_p} \mathcal{L}(\mathbf{g}_p). \tag{3.17}$$

The position with maximum amplitude $\hat{\mathbf{s}}$ corresponds to the estimated position of the strongest acoustic source. For reasons of efficiency a relatively wide spacing between the grid points $\mathbf{g}_p$, which are distributed throughout the volume of interest, of typically $0.2 \ldots 0.3$ m, is used.[1] To increase the accuracy of the source localization process, a finer grid is then defined in a volume of $\pm 0.3$ m in $\mathrm{x}, \mathrm{y}, \mathrm{z}$ direction centered around the found position of the maximum $\hat{\mathbf{s}}$, with a step size of $\Delta\mathrm{x} = \Delta\mathrm{y} = \Delta\mathrm{z} = 0.05$ m. Analog to the acoustic map $\mathcal{L}(\mathbf{g}_p)$, a secondary map $\mathcal{L}(\mathbf{g}_{p'})$ is generated defined over the finer grid points $\mathbf{g}_{p'}$, where $p' = 1, \ldots, P'$. By again picking the position corresponding to the maximum peak of the secondary acoustic map we obtain a more accurate estimate $\hat{\mathbf{s}}'$ of the true source location.[2] The found position on the finer grid

$$\hat{\mathbf{s}}' = \arg \max_{\mathbf{g}_{p'}} \mathcal{L}(\mathbf{g}_{p'}), \tag{3.18}$$

is then stored as the estimated source position.

### 3.1.3 Acoustic Map De-Emphasis

After the strongest peak is found simply searching for the $2^{nd}, 3^{rd}, \ldots$ peak can lead to wrong position estimates due to ghost peaks that can be higher in amplitude than the peak of an actual source position. Besides the sparsity weighting $\mathcal{P}(\tilde{\mathbf{b}}_p)$ another approach similar to the one presented in [BOS10, BOS08] is used to further reduce the appearance of undesired ghost peaks in the acoustic map, which is called acoustic

[1] The spacing should not be too big since peaks might then fall through the grid raster and might not be captured at all.

[2] A more accurate estimate $\hat{\mathbf{s}}'$ of the true source position $\mathbf{s}$ is also beneficial for the following acoustic map de-emphasis procedure, since peaks in the histograms $\tilde{\boldsymbol{h}}^m$ can be removed more precisely (see Ch. 3.1.3).

map de-emphasis. Here after the strongest peak of the secondary acoustic map is found, the corresponding angular regions in all histograms $\boldsymbol{h}^m$, $m = 1, \ldots, M$ are attenuated by multiplication with a window function. We use a 2-dimensional Gaussian window.

Let us assume that $\hat{\mathbf{s}}'$ is the obtained position of the strongest peak. With Eq. 3.2, $\boldsymbol{\Phi}_{p'_{max},m}$ expresses the direction of position $\hat{\mathbf{s}}'$ from the $m^{th}$ arrays point of view in the array's local coordinate system, where $p'_{max}$ is the index of $\hat{\mathbf{s}}'$ on the fine grid.

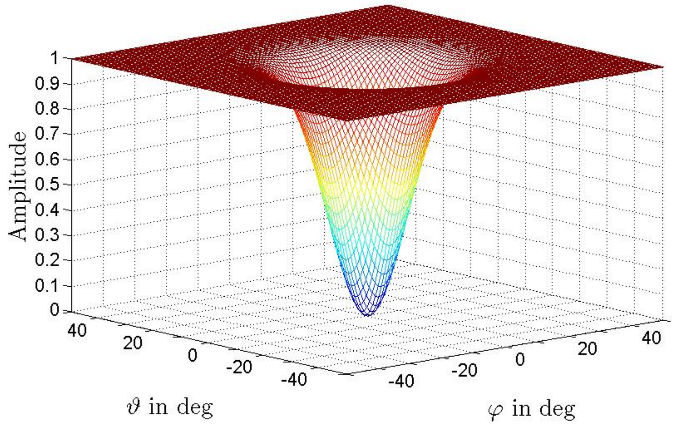



Figure 3.6: Window function $\mathcal{W}(\varphi, \vartheta)$

Note that instead of $p$, $p'$ is now used since we work on the finer secondary grid. For each array we can determine the code book direction $\boldsymbol{\theta}_{\check{q}_{p'_{max},m}}$ with index $\check{q}_{p'_{max},m}$ from Eq. 3.4 that is closest to $\boldsymbol{\Phi}_{p'_{max},m}$. A window function $\mathcal{W}$ defined as

$$\mathcal{W}(\varphi, \vartheta) = \left(1 - e^{-\frac{(\varphi^2+\vartheta^2)}{2\sigma^2}}\right), \tag{3.19}$$

and illustrated in Fig. 3.6, is shifted to the corresponding direction $\boldsymbol{\theta}_{\check{q}_{p'_{max},m}}$ by circular convolution (denoted by $\circledast$) with a shifted dirac delta pulse

$$\mathcal{W}_m(\varphi, \vartheta) = \mathcal{W} \circledast \delta(\boldsymbol{\theta}_{\check{q}_{p'_{max},m}}), \tag{3.20}$$

which yields $M$ shifted window functions $\mathcal{W}_m$ that are centered around the directions $\boldsymbol{\Phi}_{p'_{max},m}$ of $\hat{\mathbf{s}}'$ seen from the arrays point of view. We then evaluate the $m^{th}$ shifted window function at the code book directions $\mathcal{W}_m(\boldsymbol{\theta}_q)$, $q = 1, \ldots, Q$, which can be written in vector notation as $\boldsymbol{w}_m = [\mathcal{W}_m(\boldsymbol{\theta}_1),\ \mathcal{W}_m(\boldsymbol{\theta}_2), \ldots,\ \mathcal{W}_m(\boldsymbol{\theta}_Q)]^T$. Multiplication of $\boldsymbol{w}_m$ with the corresponding histograms $\tilde{\boldsymbol{h}}^m = [h_1,\ h_2, \ldots,\ h_Q]^T$ in Eq. 3.21 attenuates the angular regions in each histogram $\tilde{\boldsymbol{h}}_m$ that correspond to source position $\hat{\mathbf{s}}'$ and yields a new set of histograms $\tilde{\boldsymbol{h}}^m_{(1)}$ with

$$\tilde{\boldsymbol{h}}^m_{(1)} = \mathrm{diag}\{\boldsymbol{w}_m\}\tilde{\boldsymbol{h}}^m. \tag{3.21}$$

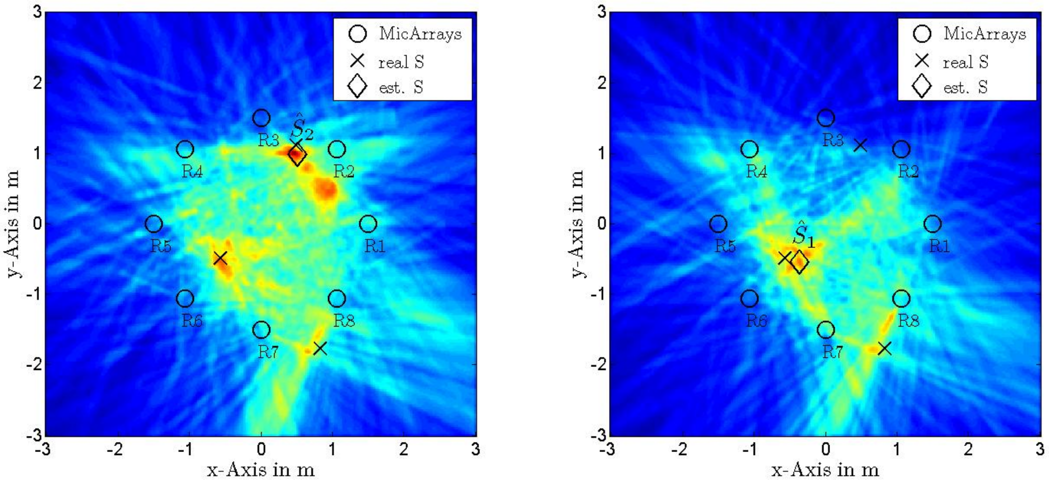
A new acoustic map $\mathcal{L}_{(1)}(\cdot)$ is then computed from the modified histograms $\tilde{\boldsymbol{h}}^m_{(1)}$, in the same way that $\mathcal{L}(\cdot)$ is computed from $\tilde{\boldsymbol{h}}^m$, to find the position of the second source, which corresponds to the global maximum of $\mathcal{L}_{(1)}(\cdot)$. The two step process of peak-picking followed by acoustic map de-emphasis is repeated until the amplitude of the found peak drops below a certain threshold $\mathcal{L}_{Th}$, which depends on the acoustic environment and has to be found empirically. The basic scheme of the localization algorithm is illustrated in Fig. 3.9. In a real world application the scheme in Fig. 3.9 is executed for each time frame. The peak amplitudes in the acoustic maps can vary from frame to frame, hence it is difficult to find a common threshold $\mathcal{L}_{Th}$ for all frames. Therefor every acoustic map is normalized to the value of its maximum peak as

$$\mathcal{L}(\mathbf{g}_p) = \frac{\mathcal{L}(\mathbf{g}_p)}{\max(\mathcal{L}(\mathbf{g}_p))}, \tag{3.22}$$

and for a threshold of e.g. $\mathcal{L}_{Th} = 0.3$, the subsequent steps of peak picking and acoustic map de-emphasis is done as long as the peak $\hat{\mathbf{s}}'_i$ of $\mathcal{L}(\mathbf{g}_p)$ after acoustic map de-emphasis is greater than $\mathcal{L}_{Th} \cdot \mathcal{L}(\hat{\mathbf{s}}'_1)$, where $\mathcal{L}(\hat{\mathbf{s}}'_1)$ denotes the amplitude of the strongest peak.

The whole localization procedure is illustrated in Fig. 3.9, 3.7 and 3.8. Fig. 3.7a shows the acoustic map $\mathcal{L}(\cdot)$ from a measurement that was conducted with a total number of 8 tetrahedral microphone arrays positioned on a circle with radius $R = 1.5$ m and 3 active sources. The sources and microphone arrays were positioned approximately in the same 2-dimensional plane parallel to the x-y plane with a height of 1.5 m. The illustrated acoustic maps are the acoustic maps corresponding to that plane. Array positions $\mathbf{r}_m$ ($\circ$) as well as true source locations $\mathbf{s}$ ($\times$) and estimated source locations $\hat{\mathbf{s}}'$ ($\diamond$) are also illustrated in Fig. 3.7.

The strongest peak that is found in the acoustic map $\mathcal{L}(\cdot)$ by the localization procedure described above is the one generated by source $S_2$. The estimated source location $\hat{\mathbf{s}}'_2$ yields a good result, as it lies only slightly off the true source position $\mathbf{s}_2$. After $\hat{\mathbf{s}}'_2$ is stored as the first source location, directions $\boldsymbol{\theta}_{\check{q}_{p'_{max},m}}$ are calculated for all microphone arrays and with Eq. 3.4 the angular regions around $\boldsymbol{\theta}_{\check{q}_{p'_{max},m}}$ in the histograms $\tilde{\boldsymbol{h}}^m$ of each array are attenuated (de-emphasis) by multiplication with the




(a) acoustic map $\mathcal{L}$

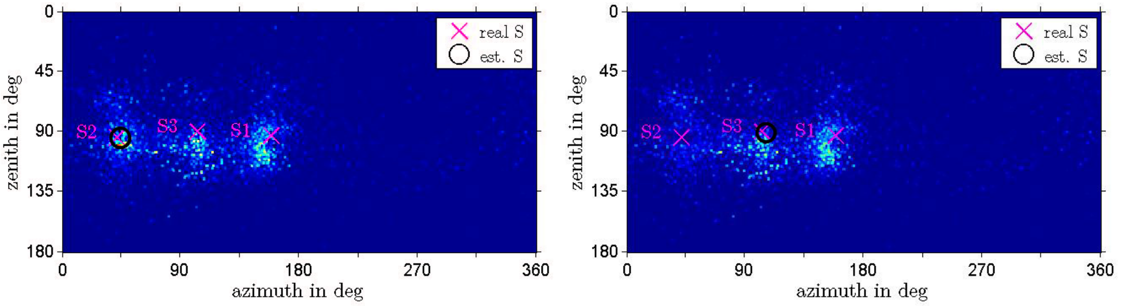
(b) acoustic map $\mathcal{L}_{(1)}$

Figure 3.7: 2-dimensional acoustic map of an acoustic scene with 8 arrays ($\circ$) and 3 sources ($\times$), before $\mathcal{L}(\cdot)$ and after $\mathcal{L}_{(1)}(\cdot)$ acoustic map de-emphasis. Estimated source positions are indicated by $\diamond$.

window functions $\mathcal{W}_m$ as described in Eq. 3.20. Fig. 3.8 shows the de-emphasis of the histograms exemplarily for array 1. Comparing the histograms before ($\tilde{\boldsymbol{h}}^1$) and after ($\tilde{\boldsymbol{h}}^1_{(1)}$) the de-emphasis, it can be seen that the histogram entries around the estimated direction $\hat{\mathbf{s}}'_2$ ($\circ$) of the source $S_2$ have been attenuated. The new acoustic map $\mathcal{L}_{(1)}(\cdot)$ that is generated from the attenuated histograms $\tilde{\boldsymbol{h}}^m_{(1)}$ is shown in Fig. 3.7b. Note that since we de-emphasized the histograms and not the acoustic map directly, not only the main peak caused by source $S_2$, but also all probability contributions in the entire acoustic map that were caused by $S_2$ have vanished. That also includes the ghost peak that otherwise would have been falsely picked as the location of the second strongest source, since it is higher in amplitude than the peak corresponding to the actual second strongest source which is $S_1$.

The example illustrated in Fig. 3.7a and Fig. 3.7b shows the importance of the acoustic map de-emphasis procedure to increase the robustness against ghost peaks.

Finding a suitable value of $\sigma$ for the window function $\mathcal{W}(\varphi, \vartheta)$ from Eq. 3.19 that is used in the de-emphasis process is a trade off. Small values of $\sigma$ lead to narrow windows, which only attenuate the histograms in close vicinity to the peak positions. This is beneficial if multiple peaks are positioned close to each other in the histograms, since a narrow window only attenuates the peak that it corresponds to and does not influence the other peaks drastically. On the other hand, as can be seen in Fig. 3.8, peaks can be different in shape, i.e. since some peaks are wider than others, the




(a) Histogram $\boldsymbol{h}^1$

(b) Histogram $\boldsymbol{h}^1_{(1)}$

Figure 3.8: Histogram of microphone array 1 (a) before and (b) after acoustic map de-emphasis, from an acoustic scene with 8 arrays and 3 sources.

angular region that has to be attenuated varies for each peak. Hence if we choose a narrow window with small $\sigma$ it might not cover the whole peak and parts of the peak can remain in the histograms, which might again lead to erroneous ghost peaks in the acoustic map. Further Fig. 3.8a shows that the estimated DOAs calculated from the $\boldsymbol{h}^m$ of all tetrahedral arrays do not always match up perfectly with the exact peak directions of each individual array. A fact that also endorses the use of wider window functions $\mathcal{W}(\varphi, \vartheta)$, i.e. larger values for $\sigma$. The ideal window $\mathcal{W}(\varphi, \vartheta)$ would be one that adapts to the shape of the corresponding peak, but that goes beyond the scope of this work. The value of $\sigma$ was found empirically as a balance between good separation between peaks and good coverage of the peak by the window function $\mathcal{W}(\varphi, \vartheta)$.

The computational complexity of the acoustic map algorithm depends on the number of grid points $\mathbf{g}_p$, i.e. on the size of the observed volume and on the grid resolution. A way to reduce the computational complexity could be to further improve the concept of different grid resolutions from Ch. 3.1.2. That is, to start with a really coarse grid and to iteratively close in on the true source position by increasing the grid resolution in a confined volume around the found maximum, where the size of the rastered volume gets smaller every iteration. To prevent the projections into space of the array's histograms from falling through the grid raster in the case of a coarse grid, the histograms have to be filtered with wide Gaussian windows.

## 3.2 Linear Intersection Algorithm

In this chapter the second localization algorithm referred to as linear intersection algorithm will be presented. The algorithm is based on a method that was initially

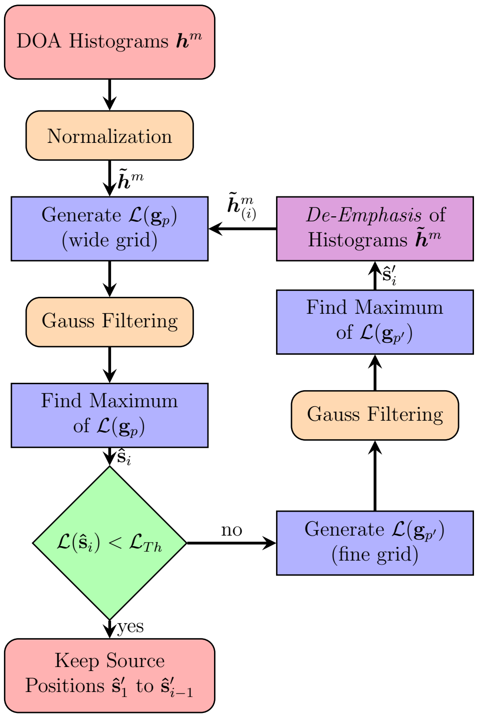



Figure 3.9: Basic scheme of the acoustic map localization algorithm.

introduced by Brandstein et. al. in [BAS97, BS97] as a closed form location estimator for TDOA based acoustic source localization. The basic idea is to use the DOA estimates of multiple distributed microphone arrays as direction vectors of bearing lines that extend from the array positions into space. The point, where the lines intersect or come closest to each other denotes the source location. In the following, the algorithm will be described in detail.

### 3.2.1 Intermediate Points

In analogy to Ch. 3.1 we assume that $M$ tetrahedral arrays are distributed in 3-dimensional space and their positions are denoted by $\mathbf{r}_m$, $m = 1, \ldots, M$ in a global cartesian coordinate system. For each tetrahedral array a local cartesian coordinate system is defined to describe possible rotations of the array orientations. The transformation between local and global coordinate system is done by applying an array dependent rotation matrix $\mathbf{R}_m$ as in Eq. 3.3. Let $\hat{\boldsymbol{\theta}}'_{\mathrm{k},m}$ (see Eq. 2.2, Ch. 2.1) be the DOA estimate of the $m^{th}$ array at frequency index k, where $\mathrm{k} = 1, \ldots, \mathrm{K}$ are the frequency indices obtained by the DFT of a single data frame. $\hat{\boldsymbol{\theta}}'_{\mathrm{k},m}$ describes the DOA estimate in the local coordinate system of the $m^{th}$ tetrahedral microphone array.

Considering $M$ arrays we obtain $M$ DOA estimates $\hat{\boldsymbol{\theta}}'_{\mathrm{k},m}$ at a single frequency index k. Each of the $M$ unit vectors $\hat{\boldsymbol{\theta}}'_{\mathrm{k},m}$ can be used as the direction vector of a bearing line that extends from the array position $\mathbf{r}_m$ into 3-dimensional space. Since we observe a single frequency index k for the sake of better readability the index k will be omitted. The line can be described in the local coordinate system of each array by the parametric equation [BAS97]

$$\mathbf{l}'_m = [x_j \quad y_j \quad z_j]^T = \mu_m \, \hat{\boldsymbol{\theta}}'_m, \tag{3.23}$$

where $\mu_m$ is the range of a point on the bearing line. The bearing line begins at the origin of the $m^{th}$ local coordinate system. Applying the rotation matrix $\mathbf{R}_m$, which describes the transformation from the $m^{th}$ local- to the global coordinate system, the lines $\mathbf{l}'_m$ can be expressed in the global coordinate system by [BAS97]

$$\mathbf{l}_m = \mu_m \underbrace{\mathbf{R}_m \hat{\boldsymbol{\theta}}'_m}_{\hat{\boldsymbol{\theta}}_m} + \mathbf{r}_m, \tag{3.24}$$

where $\hat{\boldsymbol{\theta}}_m = \mathbf{R}_m \hat{\boldsymbol{\theta}}'_m$ describes the transformation of the unit direction vector in local coordinates $\hat{\boldsymbol{\theta}}'_m$ to the unit direction vector in global coordinates $\hat{\boldsymbol{\theta}}_m$. From Eq. 3.24 we obtain $M$ bearing lines defined in the global coordinate system. The next step is to calculate two so called *intermediate points* from each pair of bearing lines as the points of closest intersection on the lines. The intermediate points mark points of potential source locations. For a total number of $M$ arrays $M!/(M-2)!$ intermediate points are obtained. Through the example of the two bearing lines [BAS97]

$$\mathbf{l}_j = \mu_j \hat{\boldsymbol{\theta}}_j + \mathbf{r}_j, \tag{3.25}$$

$$\mathbf{l}_v = \mu_v \hat{\boldsymbol{\theta}}_v + \mathbf{r}_v, \tag{3.26}$$

corresponding to array $j$ and $v$ and illustrated in Fig. 3.10 it will be explained in the following how the two points of closest intersection are found.

At first the shortest distance $d_{jv}$ between $\mathbf{l}_j$ and $\mathbf{l}_v$ is determined. It is measured along a line parallel to the common normal of the two lines as shown in Fig. 3.10 and given by [BAS97]

$$d_{jv} = \frac{|(\hat{\boldsymbol{\theta}}_j \times \hat{\boldsymbol{\theta}}_v) \cdot (\mathbf{r}_j - \mathbf{r}_v)|}{|\hat{\boldsymbol{\theta}}_j \times \hat{\boldsymbol{\theta}}_v|} \tag{3.27}$$

To find $\mathbf{p}_{jv}$, which denotes the intermediate point on the line $\mathbf{l}_j$ of closest intersection to the line $\mathbf{l}_v$ and the second intermediate point $\mathbf{p}_{vj}$, which is defined vice-versa,

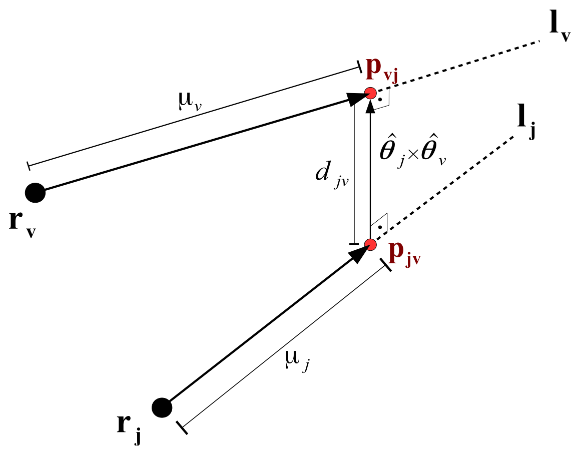



Figure 3.10: Points of closest intersection $\mathbf{p}_{jv}$ and $\mathbf{p}_{vj}$ on bearing lines $\mathbf{l}_j$ and $\mathbf{l}_v$ [BAS97]

we try to find the ranges $\mu_j$ and $\mu_v$ in Eq. 3.25, Eq. 3.26 corresponding to the two intermediate points. They fully determine a point on the bearing line.

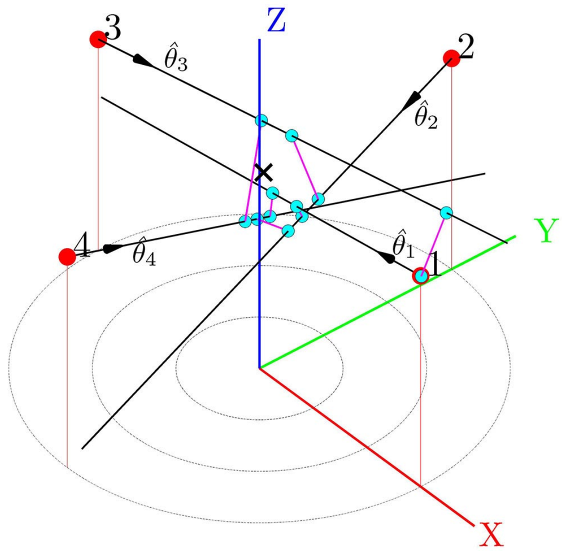



Figure 3.11: 12 points of closest intersection $\mathbf{p}_{jv}$ and $\mathbf{p}_{vj}$ (•) on bearing lines $\mathbf{l}_m$, $m = 1, \ldots, 4$, for a scenario with 4 microphone arrays (•) and a single source (×) [BAS97].

Subtracting Eq. 3.25 from Eq. 3.26 at the two points of closest intersection $\mathbf{l}_j = \mathbf{p}_{jv}$ and $\mathbf{l}_v = \mathbf{p}_{vj}$ yields

$$\mu_j \hat{\boldsymbol{\theta}}_j - \mu_v \hat{\boldsymbol{\theta}}_v = \mathbf{p}_{jv} - \mathbf{p}_{vj} + \mathbf{r}_v - \mathbf{r}_j \tag{3.28}$$

From Eq.3.27 the relation between the two intermediate points can be derived as

$$\mathbf{p}_{vj} = \mathbf{p}_{jv} + d_{jv}(\hat{\boldsymbol{\theta}}_j \times \hat{\boldsymbol{\theta}}_v). \tag{3.29}$$

Insertion into Eq. 3.28 yields the following over-determined system of linear equations

$$\mu_j \hat{\boldsymbol{\theta}}_j - \mu_v \hat{\boldsymbol{\theta}}_v = \underbrace{\mathbf{r}_v - \mathbf{r}_j - d_{jv}(\hat{\boldsymbol{\theta}}_j \times \hat{\boldsymbol{\theta}}_v)}_{\mathbf{c}} \tag{3.30}$$

$$\mathbf{A}_{jv} \boldsymbol{\mu}_{jv} = \mathbf{c}_{jv} \tag{3.31}$$

with the vector $\boldsymbol{\mu}_{jv} = (\mu_j, \quad -\mu_v)^T$ (containing two unknown scalar variables), $\mathbf{A}_{jv} = (\hat{\boldsymbol{\theta}}_j, \quad \hat{\boldsymbol{\theta}}_v)$ and the known vector $\mathbf{c}_{jv}$. Eq. 3.30 is solved in a least squares sense by using the pseudo inverse

$$\boldsymbol{\mu}_{jv} = \mathbf{A}^{+}_{jv} \mathbf{c}_{jv}. \tag{3.32}$$

Calculating $\boldsymbol{\mu}_{jv}$ in Eq. 3.32 for all pairs $j, v$ of bearing lines yields $^{M!}/_{(M-2)!}$ different range values $\mu$ and intermediate points respectively, at a single frequency index k. Since DOA estimation is done for K frequencies $f_{\mathrm{k}}$ we obtain a total number of $P = \mathrm{K}\ ^{M!}/_{(M-2)!}$ intermediate points, e.g. for $M = 4$ microphone arrays and $\mathrm{K} = 4096/2 + 1$ (4096 = DFT length) we theoretically get $P = 24588$ points. However, because of spatial aliasing DOA correction in Ch. 6 only works up to $f = 2$ kHz ($\mathrm{K} = 186$), which results in a reduced number of $P = 2232$ intermediate points for each DFT frame. Due to the fact that $^{M!}/_{(M-2)!}$ intermediate points are calculated by Eq. 3.32 and Eq. 3.25, Eq. 3.26 separately for each frequency index k, the bearing lines of all microphone arrays for frequency index k will point in the direction of the same acoustic source under the assumption of spectral disjointness (see Ch. 3.1).

### 3.2.2 Weighted k-Means Clustering

The $P$ intermediate points will be clustered around the true source positions $\mathbf{s}_i$ in 3-dimensional space, where $i = 1, \ldots, I$ and $I$ denotes the total number of active sources. For the sake of better readability the intermediate points will be denoted by $\mathbf{p}_p$, $p = 1, \ldots, P$. These point clusters are illustrated in Fig. 3.12. To make an estimate of the true source position $\mathbf{s}_i$ from the point clusters, a *weighted k-Means* clustering algorithm is used. The goal is to determine $l = 1, \ldots, L$ points $\mathbf{z}_l$, called cluster-centers and to assign each intermediate point $\mathbf{p}_p$ to some cluster $\mathbf{z}_l$ in a way that the weighted cluster energy E in Eq. 3.33 is minimized [Spa73].

$$\mathrm{E} = \sum_{p=1}^{P} \mathcal{W}_p \|\mathbf{p}_p - \mathbf{z}_l\|^2, \tag{3.33}$$

with $\mathcal{W}_p$ as the weight corresponding to intermediate point $\mathbf{p}_p$, the euclidean norm $\|.\|$ and the index $l = \mathcal{C}(p)$, where $\mathcal{C}(.)$ assigns the nearest cluster center $\mathbf{z}_l$ to the intermediate point with index $p$. Minimization of Eq. 3.33 corresponds to minimizing the sum of the weighted squared distances from each intermediate point $\mathbf{p}_p$, $p = 1, \ldots, P$ to its corresponding nearest cluster center $\mathbf{z}_l$. The cluster centers $\mathbf{z}_l$ indicate the estimated source positions. The search for $\mathbf{z}_l$ of the weighted k-Means algorithm starts with initial cluster centers that have to be handed to the algorithm and were chosen randomly in this work. We also have to tell the algorithm how many cluster centers it should determine, i.e. an estimate of the number of active sources has to be done beforehand. Ideally the number of estimated sources equals the actual number of sources, i.e. $L = I$. The estimation of active sources is implemented according to [TWH01] using the *gap statistic*, which estimates the number of clusters in a data set.

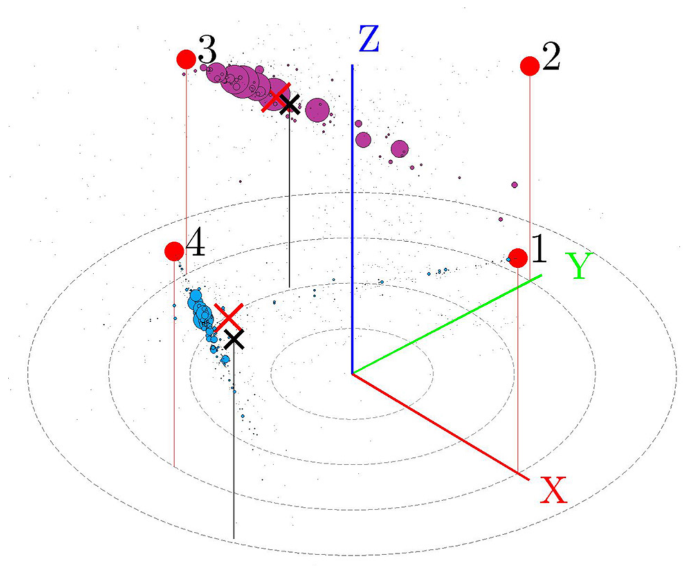



Figure 3.12: Clustered intermediate points (of 2 frames) and estimated center positions $\mathbf{z}_l$, $l = 1, 2$ ($\times$) for a scenario with 4 microphone arrays ($\bullet$) and 2 sources ($\times$).

The weighting factor $\mathcal{W}_p$ in Eq. 3.33 weights the contribution of every intermediate point $\mathbf{p}_p$ to the cluster energy E. To define $\mathcal{W}_p$ the results from Ch. 3.1 are used, in particular the acoustic map $\mathcal{L}(\cdot)$. The acoustic map indicates the probability of a present source in 3-dimensional space and is therefor well suited as a weighting factor. According to Eq. 3.16 the weighting factor is thus defined as

$$\mathcal{W}_p = \mathcal{L}(\mathbf{p}_p), \tag{3.34}$$

with intermediate point index $p = 1, \ldots, P$. The weighting criteria contained in the acoustic map are signal strength[3], distance degradation and sparsity.

### 3.2.3 Reduction of Intermediate Points

Noise and reverberation in the microphone signals can lead to significant errors in DOA estimates and intermediate point positions. Apart from emphasizing reliable and de-emphasizing unreliable intermediate points through weighting, as a second step the set of intermediate points is reduced according to different criteria before it is handed to the weighted k-Means clustering procedure. As a result the clusters of intermediate points become more distinct and lead to center positions, which estimate the true source positions more accurately. In this chapter, the different criteria will be described.[4]

**Range Criterion:** Solving Eq.3.32 can oftentimes lead to negative range values $\mu_p < 0$[5] if the corresponding bearing lines $\mathbf{I}_j$ and $\mathbf{I}_v$ do not intersect properly due to erroneous DOA estimates. Since we are only interested in positive range values $\mu_p \geq 0$, all intermediate points $\mathbf{p}_p$ with a corresponding $\mu_p < 0$ are removed from the total set of intermediate points.

**Angle Criterion:** Let us again consider the scenario from Fig. 3.11 with 4 microphone arrays positioned on a circle with radius $r_0$ in azimuth steps of $\Delta\varphi = 90°$ and a single source positioned in the circle's center at the same height as the microphone arrays. Fig. 3.13a shows the intermediate points generated by microphone array pairs $(1, 2)$, $(1, 3)$ and $(1, 4)$. Whereas array pairs $(1, 2)$ and $(1, 4)$ generate accurate intermediate points, the points generated by array pair $(1, 3)$ exhibit a great variance in the direction of the connecting line between the two arrays. This is due to the flat

[3]The acoustic map is generated from the DOA histograms $\boldsymbol{h}^m$, which are weighted by the signal strength factor $g_{\mathrm{K}}$ in Eq. 2.3.

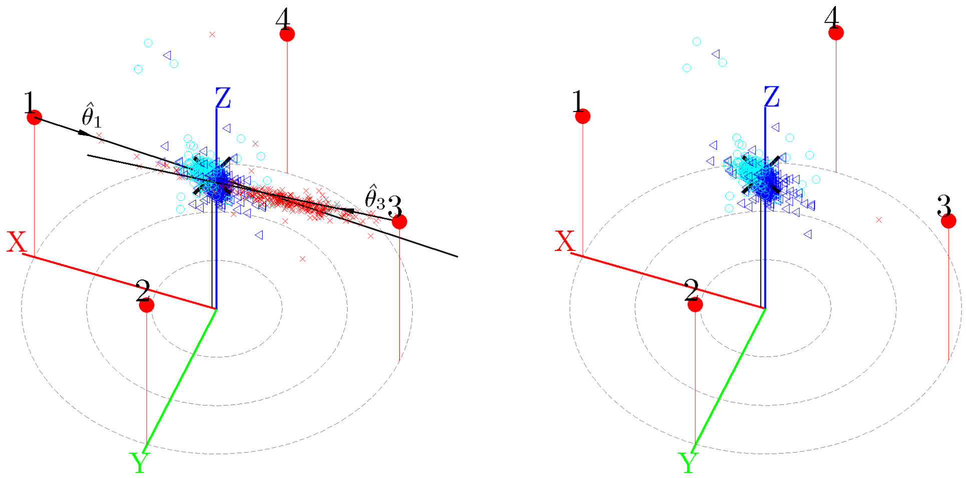
[4]Reducing the number of intermediate points also reduces the computational cost of the algorithm.

[5]As already done for the intermediate points, the index $jv$ is replaced by $p$ and $\mu_p$ denotes the range that corresponds to intermediate point $\mathbf{p}_p$.

intersection angles close to 180° of the bearing lines. Slight variations in the corresponding DOA estimates $\hat{\boldsymbol{\theta}}_1$, $\hat{\boldsymbol{\theta}}_3$ illustrated in Fig. 3.13a will lead to huge position errors of the generated intermediate points. The same is true for small angles. Thus all intermediate points that satisfy the following criteria are removed from the set

$$0^\circ \leq \measuredangle\{\hat{\boldsymbol{\theta}}_j, \hat{\boldsymbol{\theta}}_v\} < 30^\circ \quad \text{or} \quad 150^\circ < \measuredangle\{\hat{\boldsymbol{\theta}}_j, \hat{\boldsymbol{\theta}}_v\} \leq 180^\circ, \tag{3.35}$$

where $\measuredangle\{\hat{\boldsymbol{\theta}}_j, \hat{\boldsymbol{\theta}}_v\} = \arccos(\hat{\boldsymbol{\theta}}_j^{\mathrm{T}} \cdot \hat{\boldsymbol{\theta}}_v)$ denotes the intersection angle between the two bearing lines $\mathbf{l}_j$ and $\mathbf{l}_v$. Fig. 3.13b shows the resulting intermediate points after the angular criterion is applied. As can be seen the points generated by array pair $(1,3)$ have been removed, which yields a more accurate point cluster.




(a) Without angle criterion

(b) With angle criterion

Figure 3.13: Intermediate points generated by microphone array pairs $(1,2)$ ($\triangleleft$), $(1,3)$ ($\times$) and $(1,4)$ ($\circ$) without and with the angle criterion applied, for a scenario of 4 microphone arrays positioned on a circle and a single source ($\times$) in the center.

**Weighting Factor Criterion:** $\mathcal{W}_p$ indicates the likelihood that a source is present at intermediate point $\mathbf{p}_p$. Therefor all points $\mathbf{p}_p$ are removed from the set, that fulfill the following criterion

$$\frac{\mathcal{W}_p}{\max\limits_{p'}(\mathcal{W}_{p'})} < \mathcal{W}_{Th}, \tag{3.36}$$

where $\frac{\mathcal{W}_p}{\max\limits_{p'}(\mathcal{W}_{p'})}$ denotes the normalized weighting factor and $\mathcal{W}_{Th}$ is the threshold.

**Signal Strength Criterion:** As already mentioned, the generation of each intermediate point $\mathbf{p}_p$ involves a pair of DOA estimates $\hat{\boldsymbol{\theta}}_j(p)$, $\hat{\boldsymbol{\theta}}_v(p)$ of microphone arrays $(j, v)$. We define a normalized signal strength based measure $\mathcal{E}_p$ similar to the signal strength factor from Eq. 2.3 as

$$\mathcal{E}_p = \frac{|\mathrm{W}_{j(p)}(\mathrm{k}(p)) \cdot \mathrm{W}_{v(p)}(\mathrm{k}(p))|^{\alpha}}{\max\limits_{p'}(|\mathrm{W}_{j(p')}(\mathrm{k}(p')) \cdot \mathrm{W}_{v(p')}(\mathrm{k}(p'))|^{\alpha})}, \quad 0 \leq \mathcal{E}_p \leq 1, \tag{3.37}$$

where $\alpha = 0.5$ and $\mathrm{W}_{j(p)}(\mathrm{k}(p))$, $\mathrm{W}_{j(p)}(\mathrm{k}(p))$ denote the B-Format components W of the microphone array pair $j, v$ that generated the $p^{th}$ intermediate point $\mathbf{p}_p$.[6] I.e. high values of $\mathcal{E}_p$ mean that intermediate point $\mathbf{p}_p$ originates from DOA estimates $\hat{\boldsymbol{\theta}}_j(p)$, $\hat{\boldsymbol{\theta}}_v(p)$ that correspond to frequency indices k of high energy. Since $\mathcal{E}_p$ indicates the reliability of an intermediate point $\mathbf{p}_p$ in terms of signal strength, we set a threshold $\mathcal{E}_{Th}$ and remove the intermediate points $\mathbf{p}_p$ from the set with $\mathcal{E}_p < \mathcal{E}_{Th}$.

The whole localization procedure of the linear intersection algorithm is illustrated in Fig. 3.14.

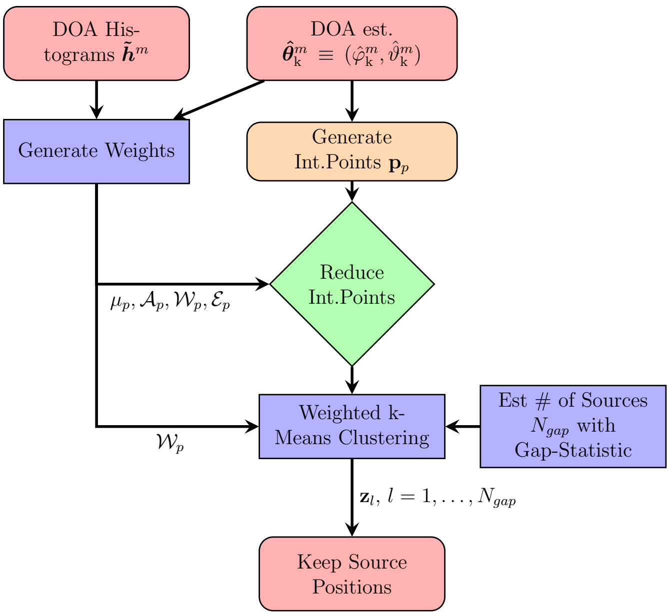



Figure 3.14: Basic scheme of the linear intersection localization algorithm.

The computational complexity of the linear intersection algorithm depends on the number of computed intermediate points $P = \mathrm{K}\ {}^{M!}/_{(M-2)!}$, which increase with the number of observed frequencies K and microphone arrays $M$.

[6] $\alpha$ is used to enhance small amplitudes of $\mathcal{E}$.

# Chapter 4

# Measuring the Tetrahedral Oktava Microphone Array

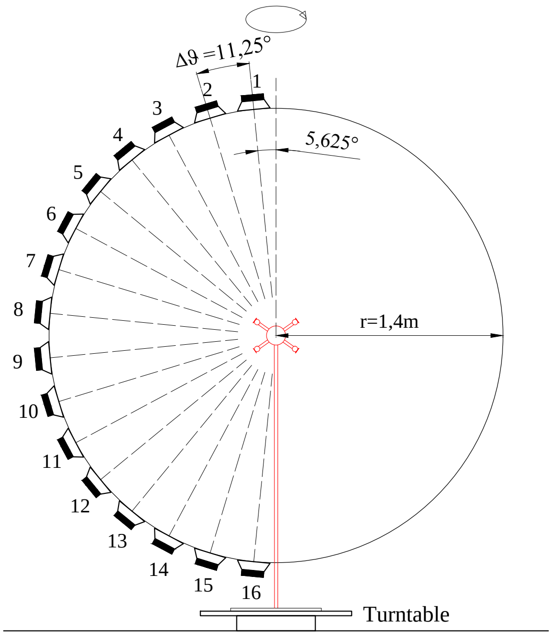



Figure 4.1: Measurement setup

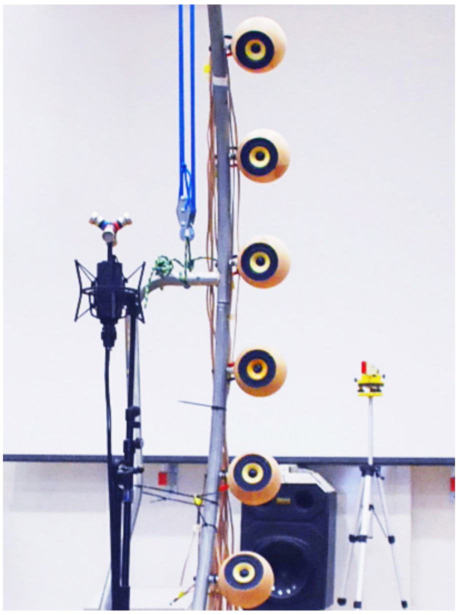


Figure 4.2: Positioning the microphone array.

This chapter explains steps to measure and analyze the frequency dependent 3-dimensional directivity patterns of the tetrahedral Oktava array's microphones. The directivity patterns are used to evaluate the quality of the microphone arrays and their applicability for DOA estimation. The microphone array is depicted in Fig. 2.1a. The measurements presented in this chapter were conducted in the IEM CUBE at the *Institute of Electronic Music and Acoustics*.

The Oktava microphone array is shown inside the measuring setup in Fig. 4.2. The measurement procedure described in the following sections was conducted for a total number of 18 Oktava 4D-Ambient microphone arrays. Measurement data discussed here refer to the 7 arrays used in the experimental study in Ch. 7, which are labeled by 1...7 here (the corresponding physical arrays carry the labels 8,13,2,4,5,6,3).

## 4.1 Measurement Setup

The transfer functions between the microphone array and one specific source direction are obtained by generating a measurement signal through a loudspeaker placed in this direction and simultaneously recording with the microphone array.
For a fine grid of many directional transfer functions the microphone array under test was placed in the center of a concentric spherical loudspeaker surrounding with a radius of $r_0 = 1.4$ m. The spatial resolution of the measurement grid was set to $\Delta\varphi = 10°$ in azimuth and $\Delta\vartheta = 11.25°$ in zenith direction (see Fig. 4.1). As the system under test is linear and time invariant (LTI), the excitation paths can be measured sequentially. We used one sixteen element semi-circular loudspeaker array (see Fig. 4.2) for measuring 16 zenith directions at every azimuth angle. To permit measuring 36 azimuth angles with one loudspeaker array, the microphone array was rotated via a computer controlled turntable in 10° steps, which leads to a surrounding virtual loudspeaker array of 16 latitude circles and 36 meridians. That amounts to $L = 576$ measured directions in total [Zau12, Pau13].

**Measurement chain:** A Computer with a running instance of *Pure Data* (PD) is connected to a RME-Madiface. The measurement signals are generated in PD, sent to the RME-M32 D/A-Converter via MADI, amplified by the Audible-Multiamp-D24 and then played back over the loudspeakers on the semicircle. Input-wise the Oktava microphone array signals are pre-amplified and A/D converted by a Presonus-DIGIMAX (ADAT output). After further ADAT/MADI conversion (RME-ADI-648) the signals are sent to the RME-Madiface which is connected to the Computer. The incoming signals are recorded in the same PD patch that is responsible for generating the output measurement signals. In addition the PD patch rotates the Computer controlled turntable by 10° after the measurement signal is sent through all 16 loudspeakers once.

**Reference Measurement:** For the reference measurement a reference microphone (omni-directional pressure transducer) was placed in the middle of the loudspeaker semicircle and an Impulse Response (IR) respectively transfer function was obtained for each of the 16 loudspeakers. The obtained transfer functions are used in Ch. 4.4.3 to compensate for the frequency response of the loudspeakers.

## 4.2 Measurement Signal

The exponentially swept sine method by Farina [Far00] was used in all of the measurements. The exponential sweeps used as measurement signals can be described mathematically in the time domain as

$$s(t) = \sin\left[f(t)\right] = \sin\left[U \cdot \left(e^{\frac{t}{V}} - 1\right)\right], \tag{4.1}$$

with the two variables

$$U = \frac{T \cdot \omega_1}{\ln\left(\frac{\omega_2}{\omega_1}\right)}, \qquad V = \frac{T}{\ln\left(\frac{\omega_2}{\omega_1}\right)}, \tag{4.2}$$

where $T$ is the sweep duration and $\omega_1, \omega_2$ denote the start and stop frequency of the exponential sweep. The instantaneous frequency $\omega(t)$ is defined as

$$\omega(t) = \frac{d\left\{f(t)\right\}}{dt} = \frac{U}{V} \cdot e^{-\frac{t}{L}}. \tag{4.3}$$

## 4.3 Obtaining the Impulse Responses

After A/D conversion ($f_s = 44.1$ kHz, 24 Bit) we obtain the recorded sweep responses $y_{lj}[n]$ from the $l^{th}$ loudspeaker to the $j^{th}$ microphone, with $l = 1, \ldots, L$, $L = 576$ and $j = 1, \ldots, 4$. To compute the desired IRs from the recorded sweep responses, either aperiodic deconvolution in time domain with a filter inverse to the excitation signal can be applied, or the aperiodic deconvolution is realized as a spectral division in the frequency domain.

$$h_{lj}\left[n\right] = IDFT\left[\frac{DFT\left(y_{lj}[n]\right)}{DFT\left(s[n]\right)}\right], \tag{4.4}$$

where $h_{lj}$ describes the IRs and $y_{lj}$ the recorded sweep response from the $l^{th}$ loudspeaker to the $j^{th}$ microphone, IDFT denotes the *Inverse Discrete Fourier Transform* and $s[n]$ denotes the discretized excitation signal from Eq. 4.1.

# 4.4 Processing the Impulse Responses

## 4.4.1 Windowing the Impulse Responses

A feature of using exponentially rising sweeps as measurement signals is that by applying the aperiodic deconvolution as shown in Eq. 4.4, all the unwanted distortion products are pushed to the left of the desired linear response and can therefore be cut out. Further, since the CUBE does not meet the requirements of an ideal anechoic chamber, adequate windowing of the IRs is performed, which allows to suppress unwanted room reflections. This is done with the first and second half of a hanning window. In the given measurement setup the floor as being the nearest room boundary to the microphone array causes the first reflection. Particularly for loudspeaker 16, which is the loudspeaker closest to the floor (see Fig. 4.1), the time delay between direct sound and first reflection impinging at the microphone array is especially short with a value of approximately 3.5 ms $\triangleq$ 154 samples (@ $f_s = 44.1$ kHz). The measured IRs were windowed accordingly to a resulting length of 150 samples, so that reflections with a path difference relative to the direct sound path of approximately 1.25 m are suppressed. In addition the floor was covered with absorptive material in close vicinity to the microphone array to reduce reflections.[1]

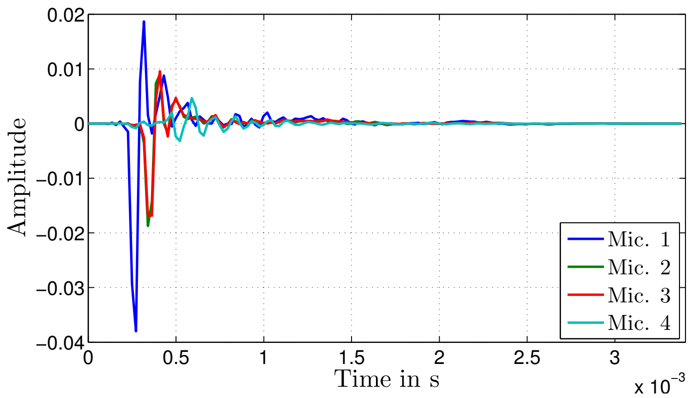



Figure 4.3: Windowed IRs of the 4 microphones with a length of 150 samples, for source position $\varphi = 0°$ and $\vartheta = 84.375°$. (Array 1)

In Fig. 4.3 the windowed IRs of the 4 microphones with a length of 150 samples are shown for a source position at $\varphi = 0°$ and $\vartheta = 84.375°$.

[1]The frequency that fits into a window of 150 Samples with one full wavelength is $f_{low} = {}^{44100\ \mathrm{Hz}}/_{150\ \mathrm{samples}} \approx 294$ Hz . It approximately denotes the lowest frequency that can be described.

### 4.4.2 Correction of Geometric Measurement Errors in Radius

To minimize phase errors that occur due to slight errors in positioning of the microphone array[2], and geometric deviations of the measurement setup[3], the measured IRs were re-aligned in time. The time alignment procedure will be described in this chapter.

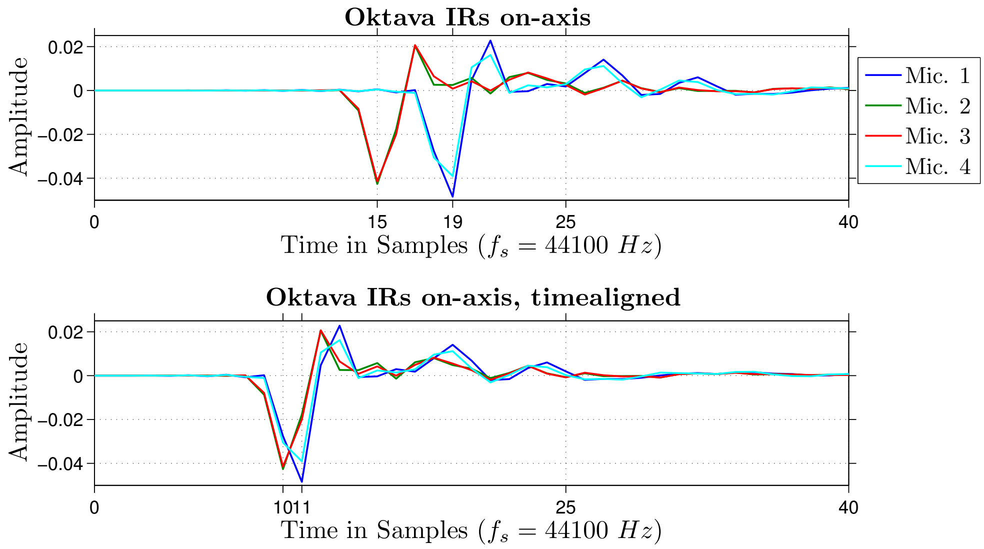



Figure 4.4: On-axis IRs before and after time alignment. (Array 1)

Time alignment is realized by using the measured Oktava IRs. As a first step, the 4 microphone IRs are summed up for each for the $L = 576$ source directions. By doing this we want to generate IRs that would be generated by a microphone that sits in the center of the microphone array. The resulting 576 IRs are then low-pass filtered (with a corner frequency of 800 Hz) to make the following „peak-picking“step easier. In this step the positions of the first minima of all 576 IRs are evaluated and for each of the 16 loudspeakers the mean value of the corresponding 36 minima positions is calculated. From the resulting 16 mean values the smallest mean value is subtracted and we obtain the mean time delay of each loudspeaker to the loudspeaker that is closest to the microphone array. The 16 delays are then used to re-align the measured IRs of the Oktava microphone array. This is done by multiplying the upsampled transfer functions $H_{lj}(\mathrm{k})$ of the Oktava microphone array with the following linear phase term

[2]The microphone array is not exactly centered.
[3]The loudspeaker semicircle deviates from an ideal semicircle.

$$\hat{H}_{lj}(k) = H_{lj}(k) \cdot e^{-i(2\pi \frac{k}{K} \cdot \text{delay}(l))}, \tag{4.5}$$

where k denotes the frequency index, K is the total number of frequency indices (DFT length) and delay($l$) describes the up-sampled delay corresponding to the $l^{th}$ loudspeaker.

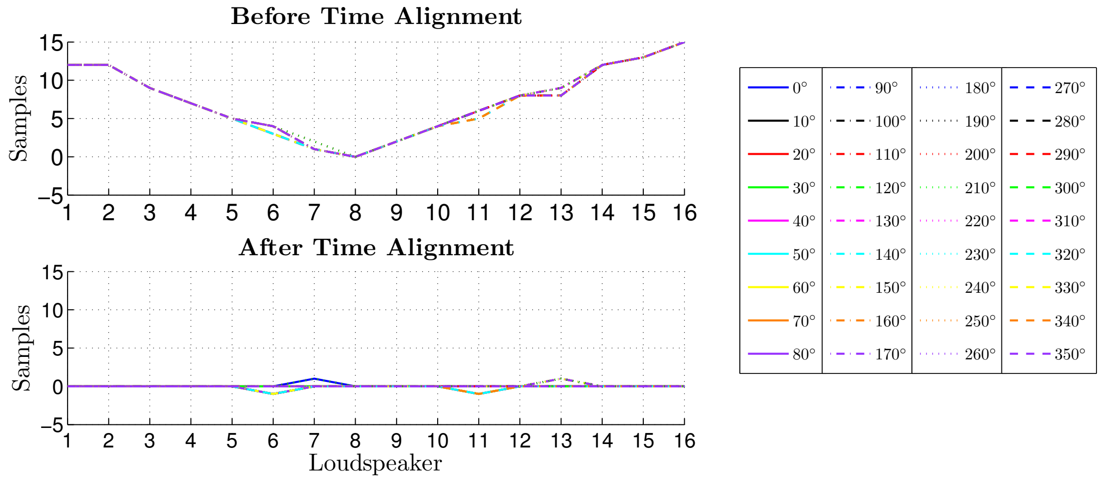



Figure 4.5: Flight time differences (in samples) in low-pass W channel for all measured Oktava IRs illustrated over the 16 loudspeakers. The legend indicates the azimuth angle. (Array 1)

Fig. 4.5 shows the loudspeaker to microphone array flight time differences of all 576 source positions, where each of the 16 loudspeaker columns contains 36 (azimuth) position values[4]. Most of the illustrated values are not visible since they are overlaid. Since the values for each loudspeaker do not vary much, it is valid to generate a mean delay value for each loudspeaker instead of applying individual delays for each source direction. Observing the first row in Fig. 4.5 it is evident, that the measurement setup exhibits a systematic error in the acoustic flight time from each loudspeaker, i.e. it is not strictly semi-circular in radius. The flight time differences are smallest for loudspeaker 8, which is positioned slightly above the equator at $\vartheta \approx 84.4°$. Moving away from the equator to the poles of the semicircle the distance increases, which means that the semicircles poles are bent away from the center of the semicircle. The second row of Fig. 4.5 shows the minima positions after the time alignment.

[4]The 36 position values of each column correspond to the chosen grid of 36 azimuth angles for each latitude circle.

The procedure shows a significant reduction of the flight time differences from a maximum of 15 samples $\hat{=}$ 11.67 cm (@ $f_s = 44100$ Hz) without time alignment to a maximum of 2 samples $\hat{=}$ 1.56 cm Samples.

### 4.4.3 Equalization of the Measurement Loudspeakers

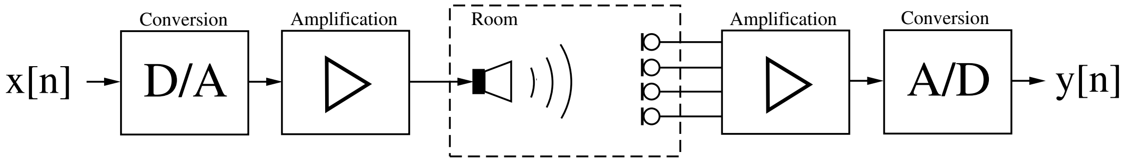



Figure 4.6: Signal chain during the measurement.

Measuring the transfer functions between the different source positions and the 4 microphones of one Oktava microphone array, apart from the desired microphone characteristics also characteristics of the room, which are minimized by windowing as described in Ch. 4.4.1, the D/A and A/D converters, the microphone pre-amplifier and most prominent the loudspeakers are captured as shown in Fig. 4.6. As the influences of conversion and pre-amplification are negligible, they are not included in Eq. 4.6, which shows the Fourier transform of the measured signal as

$$Y(f) = X(f) \cdot H_{LS}(f) \cdot H_{Mic}(f). \tag{4.6}$$

To minimize these undesired influences equalization based on the reference measurement is performed on the measured Oktava IRs. An equalization filter is generated from each of the 16 transfer functions of the reference measurement. The equalization filters are designed as minimum-phase filters[5] as in [PL06] and their magnitude responses are limited to a dynamic range of 20 dB [Zau12].

Fig. 4.7 shows the frequency responses of the minimum-phase filters. For the sake of better readability only the frequency responses corresponding to even loudspeaker numbers are shown.

Further in Fig. 4.8 the magnitude responses of microphone 1...4 are illustrated for one source position before and after the equalization procedure.

[5] Minimum-phase filters are used because we do not want the filters to have linear phase components. If the filters had linear phase components, then the time alignment performed in Ch. 4.4.2 would be destroyed.

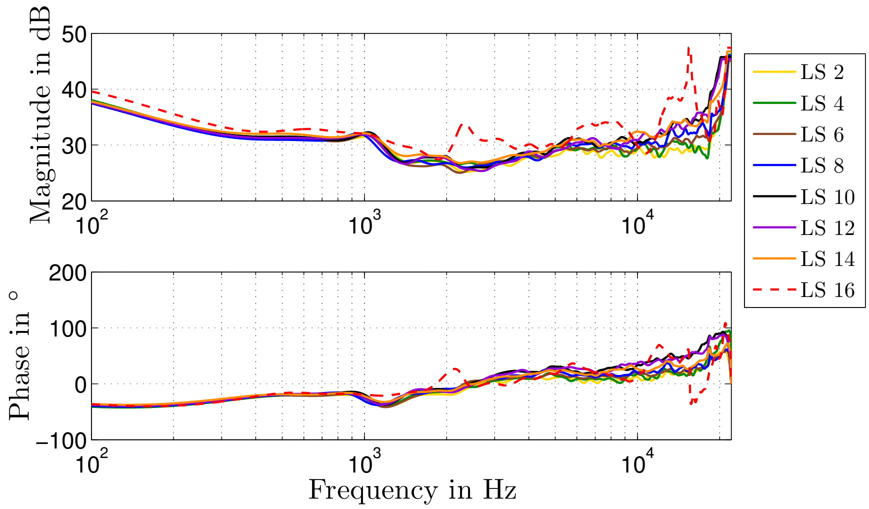
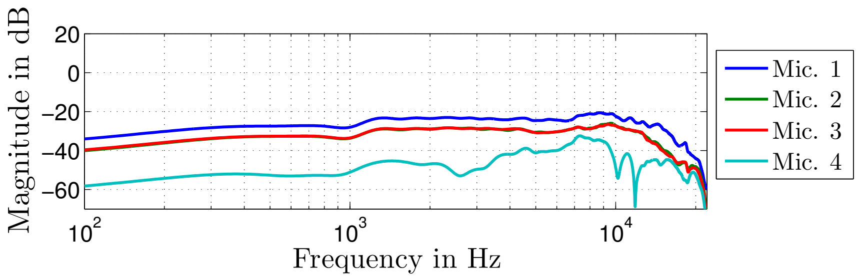
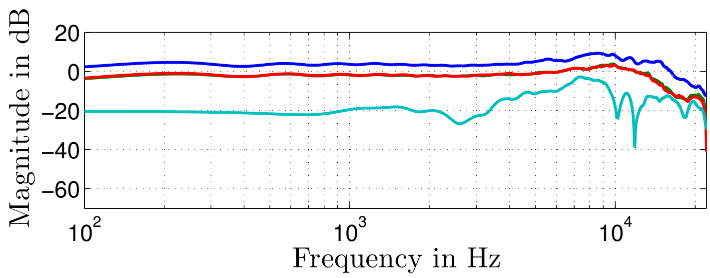



Figure 4.7: Loudspeaker equalization filters (miminum-phase).




(a) Before loudspeaker equalization




(b) After loudspeaker equalization

Figure 4.8: Magnitude response before and after loudspeaker equalization of the 4 microphones for source position $\varphi = 0°$ and $\vartheta = 84.375°$. (Array 1)

### 4.4.4 Improved Geometry Estimation and Gain Correction for the Arrays

In reality the on-axis orientations of the 4 microphones of an Oktava microphone array deviate from the ideal on-axis microphone orientations given in Tab. 2.1. One reasons for this is, that the construction of the Oktava microphone arrays does not seem to be fully symmetric, i.e. it varies from an ideal tetrahedral design. Also the effective orientation of the microphones changes over frequency. Applying A- to B-Format conversion later on, it is crucial to utilize the true on-axis orientations when generating the conversion matrix from Eq. 2.8, otherwise DOA estimation errors are induced.

This chapter therefore deals with the estimation of individual on-axis directions of the 4 microphones. To get an estimation of the on-axis direction of one microphone, the directivity pattern of that microphone is investigated. We can define a vector that contains the discrete directional response of the $j^{th}$ microphone as [Zau12]

$$\boldsymbol{h}_j(f) = \begin{pmatrix} h_j(\boldsymbol{\theta}_1, f) \\ h_j(\boldsymbol{\theta}_2, f) \\ \vdots \\ h_j(\boldsymbol{\theta}_L, f) \end{pmatrix}, \tag{4.7}$$

where $\boldsymbol{\theta}_l$ denotes the direction of the $l^{th}$ loudspeaker with $l = 1, \ldots, L$.

If the magnitude of $h_j(\boldsymbol{\theta}_l, f)$ is plotted over all loudspeaker directions with index $l = 1, \ldots, L$ in 3-dimensional space, the result is a directivity pattern that shows the direction dependent sensitivity of one microphone at the chosen frequency $f$. We define the on-axis direction at frequency $f$ as the direction with the highest amplitude. However, the observed grid with its measured $L = 576$ loudspeaker directions is too coarse to make accurate estimations of the on-axis directions. We therefore change to a finer grid step size with an angular resolution in $\theta$ and $\phi$ of $1°$, which makes $L = 180 \times 360 = 64800$ grid points respectively source directions. Unlike the 576 source directions the 64800 source directions have not been measured directly. We use spherical interpolation to interpolate the 64800 coefficients from the given 576 measured fourier coefficients.

The interpolated directivity patterns are shown in Fig. 4.9. The mathematical background for the spherical interpolation is presented in Ch. 4.5.1. To account for the fact, that the on-axis direction is frequency-dependent, we average over all on-axis directions from 100 . . . 2000 Hz, which yields a mean on-axis direction that is used for the A- to B-Format encoder later on. The found mean on-axis orientations for microphone array 1 are given in Tab. 4.1 and illustrated in Fig. 4.9.

One thing that is noticeable when looking at the $\vartheta$ values of Tab. 4.1 and the on-axis zenith angle distribution of all used arrays illustrated in Fig. 4.10 is, that they are

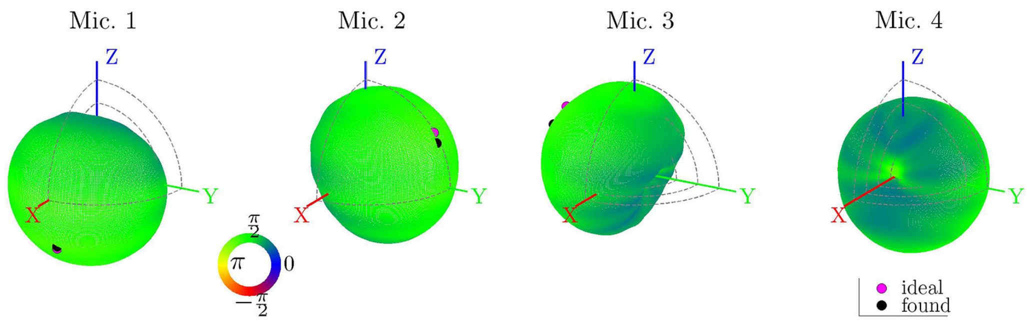



Figure 4.9: 3D Directivity patterns, ideal on-axis orientations (●) and found on-axis orientations (●) with fine grid (64800 Points) for all 4 microphones at a frequency of $f = 1000$ Hz. (Array 1)

not symmetrical around the equator ($\vartheta = 90°$) in contrast to the ideal values. This might be a characteristic of the microphone array or due to measurement errors, e.g. a positioning offset of the microphone array from the semicircles center or because the semicircle itself is not symmetric around the equator axis. Further it can be observed, that the on-axis zenith angles of all microphone arrays lie closer to the equator than the ideal values. A reason for this might be the design of the microphone arrays. The microphone orientations seem to be slightly squeezed towards the equator.[6] By using the found microphone orientations from Tab. 4.1 in the A- to B-Format conversion, the effects of these measurement errors are reduced.[7]

| Mic. | $\varphi$ in ° | $\vartheta$ in ° |
|---|---|---|
| 1 | 1 | 123 |
| 2 | 87 | 61 |
| 3 | 276 | 67 |
| 4 | 180 | 116 |

Table 4.1: On-axis dir., fine grid. (Ar. 1)

| Mic. | $\varphi$ in ° | $\vartheta$ in ° |
|---|---|---|
| 1 | 0 | 125.26 |
| 2 | 90 | 54.74 |
| 3 | 270 | 54.74 |
| 4 | 180 | 125.26 |

Table 4.2: On-axis dir., ideal.

The last step of processing the measured Oktava IRs is gain correction. Let us assume that one microphone has a larger gain compared to the other three microphones. What happens during DOA estimation is that the estimated directions of sources are drawn to that microphones direction. The first row of Fig. 4.11 shows the on-axis amplitude responses of the four microphones, which ideally should be matched in the

[6] A detailed construction plan was requested from Oktava, but could not be obtained.

[7] At the beginning of this chapter it was also mentioned, that the microphone on-axis orientations change over frequency. To reduce the consequences of this undesired behavior, one could also calculate mean on-axis orientations over multiple frequencies, which is not done in this work.

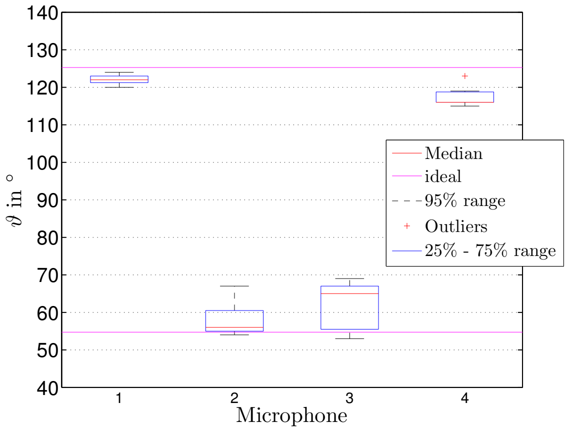
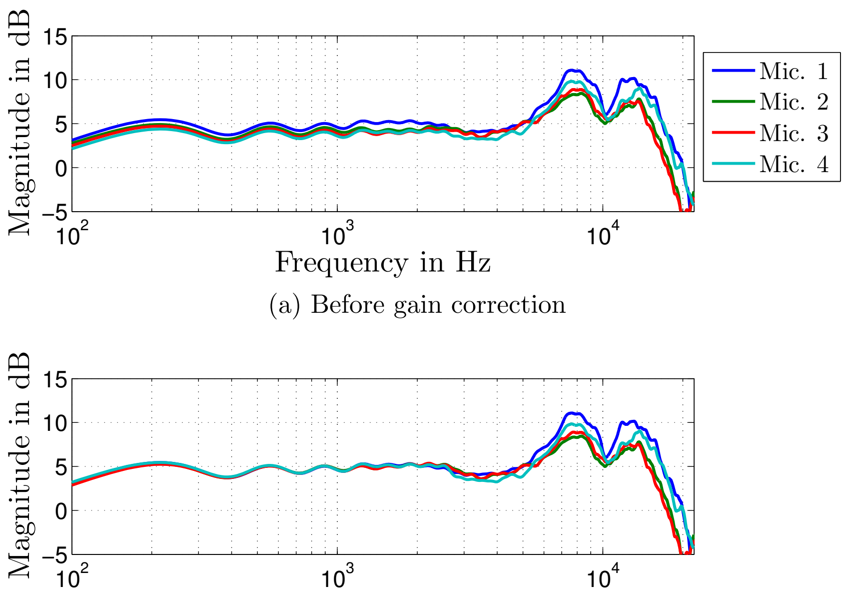



Figure 4.10: Distribution of on-axis zenith angles $\vartheta$ in ° for microphone arrays R = {1,2,3,4,5,7,8}.

relevant frequency range from approximately 40 . . . 2000 Hz.[8]




(a) Before gain correction

(b) After gain correction

Figure 4.11: On-axis amplitude response before and after gain correction. (Array 1)

In Fig. 4.11a however microphone 1 exhibits a slightly higher gain compared to the others. As a way of compensation, gain correction factors are calculated from the on-axis amplitude responses for each frequency index up to 2 kHz and are then applied to the microphone signals to match the amplitude responses of the 4 microphones up to 2 kHz (see Fig. 4.11b). Applying a IDFT, the gain-corrected IRs are obtained in

[8] The upper frequency limit of 2 kHz will be explained in Ch. 6.

time domain.

## 4.5 Characteristics of the Microphone Array

In the first part of this chapter the mathematical background needed to generate 3-dimensional directivity patterns (see Fig. 4.9) from the measured Oktava IRs is presented and these directivity patterns are analyzed in the second part of this chapter in order to evaluate the applicability of the Oktava arrays for DOA estimation.

### 4.5.1 Spherically Interpolated Directivity Patterns

To generate a spherically interpolated frequency response $x_j$ of the $j^{th}$ microphone, we want to derive a weighting vector $\boldsymbol{g} = [g_1, \ldots, g_L]^{\mathrm{T}}$ that corresponds to the driving voltages of the $L$ loudspeakers and has the same number of elements as the vector $\boldsymbol{h}$ from Eq. 4.7. We can then write

$$\mathrm{x}_j(\boldsymbol{\theta}, f) = \boldsymbol{h}_j^{\mathrm{T}}(f)\boldsymbol{g}, \tag{4.8}$$

which is equivalent to weighting and summing up the impulse responses, where $\boldsymbol{\theta}$ is the desired interpolation direction [Zau12]. To derive $\boldsymbol{g}$ let us assume that all loudspeakers are ideal point sources weighted by their corresponding weighting factor $g_l$, $l = 1, \ldots, L$. Then we can represent them as a continuous driving distribution of weighted and summed Dirac delta functions [9]

$$f(\boldsymbol{\theta}) = \sum_{l=1}^{L} \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_l) g_l. \tag{4.9}$$

By decomposing $f(\boldsymbol{\theta})$ into spherical harmonics we get the following modal source strength coefficients [ZPF09]

$$\hat{\phi}_n^m = \sum_{l=1}^{L} Y_n^m(\boldsymbol{\theta}_l) g_l, \tag{4.10}$$

and if we assume, that the system is band-limited to a maximum order $N_C$ we can write Eq. 4.10 in matrix vector notation.

[9] A point source has a Dirac delta function as its driving distribution. This is clear when we insert a Dirac delta on the right side of the Helmholtz equation (perturbation function). The Helmholtz equation becomes in-homogeneous and it is solved by the Greens function $\boldsymbol{G}$, which describes the propagation of a point source.

$$\hat{\boldsymbol{\phi}}_{N_C} = \mathbf{D}_{N_C}\boldsymbol{g}, \tag{4.11}$$

with $\mathbf{D}_{N_C} = \big(\mathbf{y}_{N_C}(\boldsymbol{\theta}_1), \mathbf{y}_{N_C}(\boldsymbol{\theta}_2),\ \ldots, \mathbf{y}_{N_C}(\boldsymbol{\theta}_L)\big)$. The next step is to match the interpolated modal source strength $\hat{\boldsymbol{\phi}}_{N_C}$ to the well known modal source strength $\boldsymbol{\phi}_{N_C}$ of a point source from the desired direction $\boldsymbol{\theta}$[10], which is characterized by $\boldsymbol{\phi}_{N_C} = \mathbf{y}_{N_C}(\boldsymbol{\theta})$. Because of the matching we can replace $\hat{\boldsymbol{\phi}}_{N_C}$ from Eq. 4.11 with $\boldsymbol{\phi}_{N_C}$ and thus $\boldsymbol{g}$ can be calculated through

$$\boldsymbol{g} = \mathbf{D}^{+}_{N_C}\boldsymbol{\phi}_{N_C} \Rightarrow \boldsymbol{g} = \mathbf{D}^{+}_{N_C}\mathbf{y}_{N_C}(\boldsymbol{\theta}), \tag{4.12}$$

where $\mathbf{D}^{+}_{N_C} = \mathbf{D}^{\mathrm{T}}_{N_C}(\mathbf{D}_{N_C}\mathbf{D}^{\mathrm{T}}_{N_C})^{-1}$ is the right-inverse of $\mathbf{D}_{N_C}$ and is called *decoder matrix*. For Eq. 4.12 to be solvable, the condition $L \geq (N_C + 1)^2$ has to be fulfilled and $\mathbf{D}_{N_C}\mathbf{D}^{\mathrm{T}}_{N_C}$ has to be well conditioned $\rightarrow \kappa(\mathbf{D}_{N_C}\mathbf{D}^{\mathrm{T}}_{N_C}) \ll \infty$ [ZPF09, Pom08].
Inserting Eq. 4.12 into Eq. 4.8 we get the resulting interpolated IR

$$\mathrm{x}_j(\boldsymbol{\theta}, f) = \mathbf{h}^{\mathrm{T}}_j(f)\mathbf{D}^{+}_{N_C}\boldsymbol{\phi}_{N_C}. \tag{4.13}$$

To determine the maximum possible order N of the *decoder matrix* $\mathbf{D}^{+}_{N_C}$ for our measuring setup with 576 loudspeaker positions, the condition number $\kappa(\mathbf{D}_{N_C}\mathbf{D}^{\mathrm{T}}_{N_C})$ has to be observed. Since we have a loudspeaker sampling grid which is similar to an equal angular resolution grid[11] we can insert a weighting matrix $\mathbf{W}$ as in [Pom08], which weights the loudspeaker nodes close to the equator more since they are more important. For the weighted and unweighted case the condition number skyrockets for $N \geq 15$, $N = 14$ is therefore the maximum spherical harmonics order we can use for interpolation.

### 4.5.2 Analysis of Directivity Patterns

To analyze the rotational symmetry of the cardioid patterns shown in Fig. 4.9, the 3-dimensional directivity patterns are rotated in a way that the on-axis direction respectively the direction of maximum sensitivity of the 3-dimensional patterns, estimated in Ch. 4.4.4, now points in the direction of the positive x-axis. In a second step slices are taken from the rotated 3-dimensional patterns in 45° steps, i.e. at $\vartheta = 0, 45, 90, 135, 180°$ as shown in Fig. 4.12 exemplarily for microphone 1.
The corresponding magnitude and phase information of the 5 slices is illustrated in Fig. 4.13 for microphone 1 and 4 at frequencies of $f = 350$ Hz and 2 kHz.

[10] The desired point source is also sitting at the loudspeaker radius.
[11] Equal Angular Resolution Grid: Equal angle resolution in longitude and latitude $\rightarrow \Delta\varphi = \Delta\vartheta$.

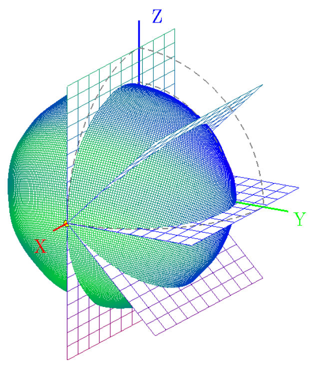



Figure 4.12: Rotated and sliced directivity pattern of microphone 1 at $f = 1500$ Hz. (Array 1)

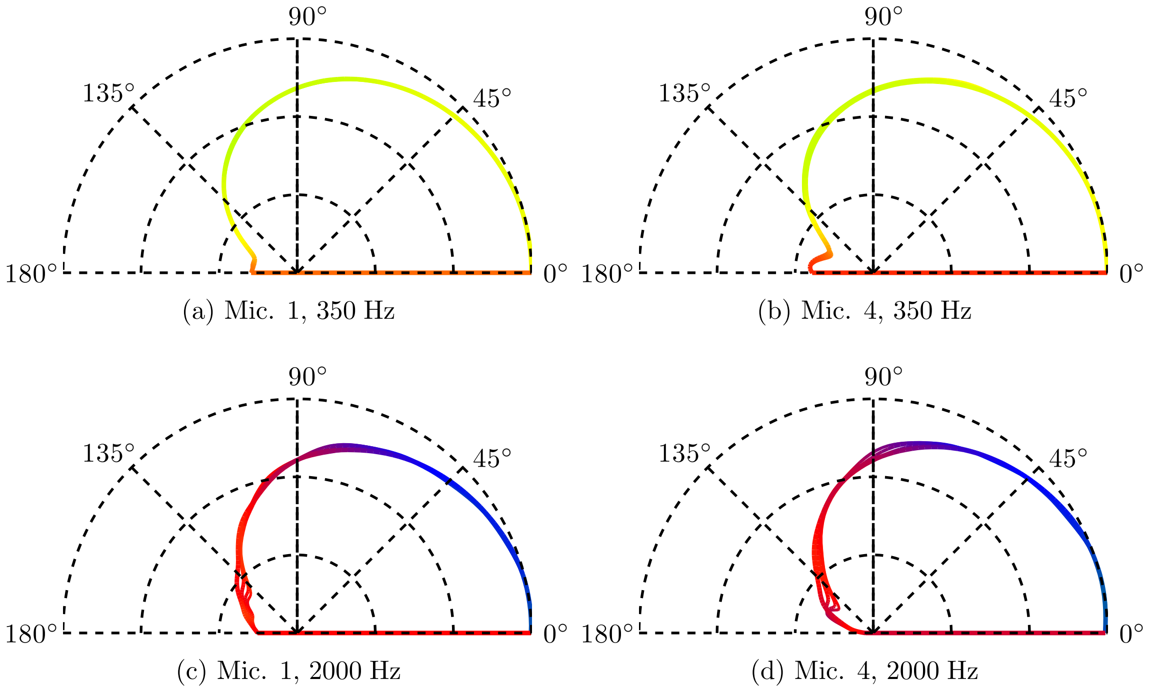
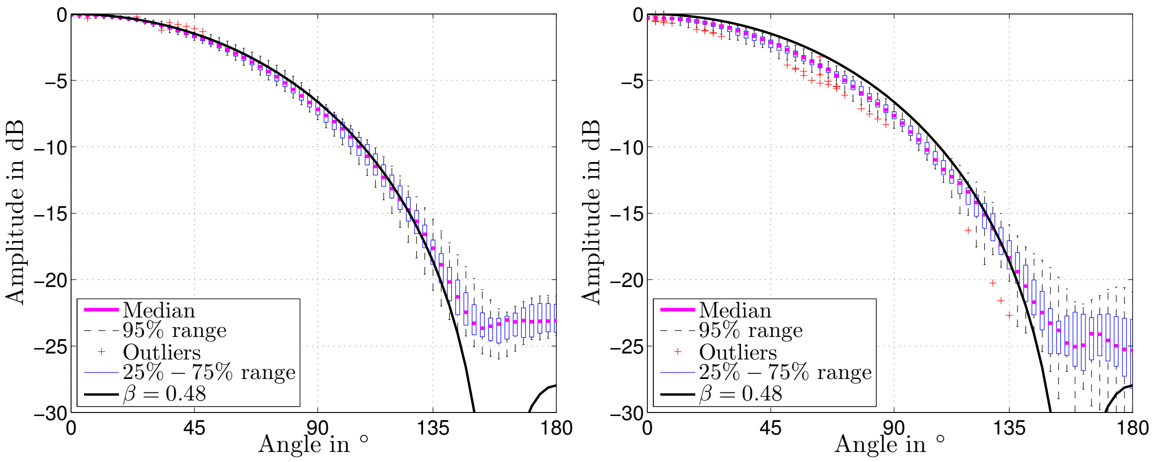



Figure 4.13: Directivity pattern slices at cutting angles $\vartheta = -90, -45, 0, 45, 90°$ for frequencies 350 Hz and 2000 Hz of Mic. 1 (left) and Mic. 4 (right). Dynamic range is 30 dB (10 dB/division). (Array 1)

At 350 Hz the microphones exhibit a standard cardioid directivity pattern with a sensitivity reduction of approximately $-7$ dB at an incident angle of 90°, whereas for a frequency of 2 kHz the shape of the directivity pattern is somewhat more narrow or more directional. In terms of rotational symmetry microphone 1...4 show similar results with relatively little variance between the slice curves for all frequencies, which indicates high symmetry. Especially for incident directions $-60°$ to $60°$. Symmetry decreases if we go from frontal (0°) to rearward incidence (180°), but is still acceptable.

Fig. 4.14 illustrates the variance of the directivity patterns between the 4 microphones. The data is obtained again by rotating and slicing the directivity patterns of all 4 microphones with cutting angles of $\vartheta = -90, -45, 0, 45, 90°$. The directivity pattern that is obtained by connecting all median values of Fig. 4.14 can be approximately described with the model from Eq. 2.9 and a factor of $\beta = 0.48$ (black curve in Fig. 4.14). The 25% to 75% range stays within 5 dB and the 95% range within 10 dB for almost all angles. The highest values are found in the area of $135 - 180°$. The variance in symmetry is a drawback of cardioid microphones and leads to errors in the DOA estimation process, as will be seen later on.




(a) All mics, 350 Hz

(b) All mics, 2000 Hz

Figure 4.14: Variance of directivity patterns for different frequencies. The data was obtained from directivity pattern slices at cutting angles $\vartheta = -90, -45, 0, 45, 90°$ from all 4 microphones. The black curve is generated from the ideal first order pattern (Eq. 2.9) with $\beta = 0.48$. (Array 1)

### 4.5.3 B-Format and Re-Diagonalization Filters

It was shown In Ch. 4.5.1 how to use the measured IRs between the loudspeakers and the 4 microphones of the microphone array to synthesize IRs that correspond to an arbitrary modal source strength excitation $\boldsymbol{\phi}_{N_C}$ at a radius of $r_0 = 1.4$ m. In this chapter we want to obtain a relation between the modal source strength $\boldsymbol{\phi}_{N_C}$ on the excitation side and the decomposed spherical harmonics coefficients $\boldsymbol{\chi}_{N_A}$ on the microphone side, where the order of excited spherical harmonics is set to $N_C = 10$

and the order of decomposed spherical harmonics is $N_A = 1$. Analog to Ch. 2.2.1.2 the relation is described by a system matrix $\dot{\mathbf{H}}_{N_A,N_C}$, which in contrast to Ch. 2.2.1.2 is based on the measured data and not on an analytical model. $N_A$ indicates the decomposition and $N_C$ the excitation order, i.e. the system matrix contains $(N_A+1)^2$ rows and $(N_C+1)^2$ columns. By means of $\dot{\mathbf{H}}_{N_A,N_C}$ we will examine the array's capability to decompose the excited spherical harmonics correctly in the array achievable subspace, i.e. for order $N_A = 1$. Also spatial aliasing caused by spherical harmonics of higher order $N_A < n \leq N_C$ will be analyzed based on $\dot{\mathbf{H}}_{N_A,N_C}$.

To derive the system matrix, we begin by re-writing Eq. 4.8. We stack its elements $x_j(f)$, $j = 1, \dots, J$ into a vector $\mathbf{x}(f) = [x_1(f) \ \dots \ x_J(f)]^{\mathrm{T}}$, where $J = 4$ is the number of microphones. This yields

$$\mathbf{x}(f) = \mathbf{H}(f)\ \mathbf{g}, \tag{4.14}$$

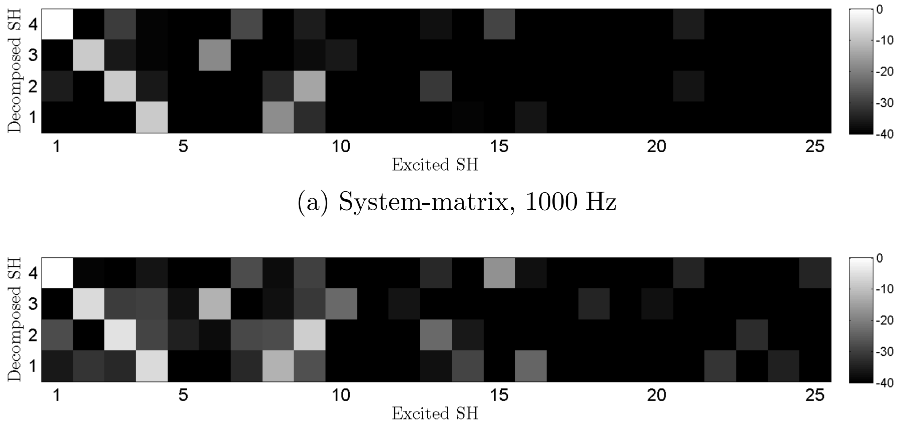
with the matrix

$$\mathbf{H}(f) = \begin{pmatrix} \mathbf{h}_1^{\mathrm{T}}(f) \\ \vdots \\ \mathbf{h}_J^{\mathrm{T}}(f) \end{pmatrix}. \tag{4.15}$$

From Eq. 4.12 it is known that the gain factors $\mathbf{g}$ can be obtained by $\mathbf{g} = \mathbf{D}_{N_C}^{+}\boldsymbol{\phi}_{N_C}(f)$. For better readability the frequency dependence is omitted from now on. Inserting $\mathbf{g}$ into Eq. 4.14 and transforming the microphone side into spherical harmonics yields the spherical wave spectrum

$$\boldsymbol{\chi}_{N_A} = \underbrace{\mathbf{Y}_{N_A}^{+}\mathbf{H}\mathbf{D}_{N_C}^{+}}_{\dot{\mathbf{H}}_{N_A,N_C}}\boldsymbol{\phi}_{N_C}, \tag{4.16}$$

where $\dot{\mathbf{H}}_{N_A,N_C}$ denotes the already mentioned system matrix. The system matrix is illustrated in Fig. 4.15 for a frequency of $f = 1000$ Hz and $f = 2000$ Hz.
Observing the first 4 columns of the system matrix (array achievable subspace) one can see that, in contrast to the system matrix of the model (see Fig. 2.7), it is not ideally diagonal. Since these erroneous elements in the system matrix are prominent for all frequencies, they lead to distorted B-Format signals $\boldsymbol{\chi}_{N_A}$ even for lower frequency ranges where spatial aliasing is small, as can be seen e.g. in Fig. 4.16b which shows a slightly tilted Z pattern. Distorted B-Format signals in turn lead to errors in the DOA estimation. Possible reasons for these erroneous elements in the system matrix




(a) System-matrix, 1000 Hz

(b) System-matrix, 2000 Hz

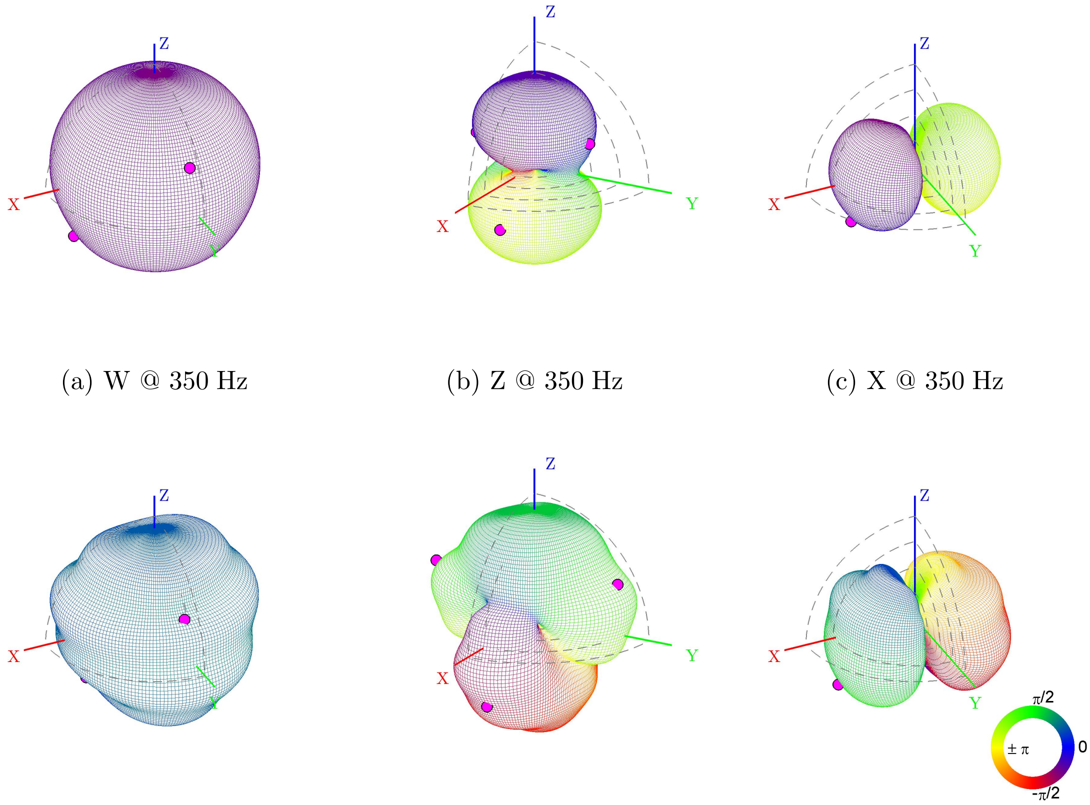
Figure 4.15: First 25 columns of system matrix $\dot{\mathbf{H}}_{N_A,N_C}$ for different frequencies. (Array 1)

might be reflections from the housing of the Oktava microphone array, mismatches in the directivity-patterns of the microphone capsules or deviations of the on-axis microphone directions from the directions obtained in Ch. 4.4.4.

Fig. 4.17 illustrates the magnitude response of each matrix element contained in the array achievable subspace of $\dot{\mathbf{H}}_{N_A,N_C}$, i.e. $\dot{\mathrm{H}}_{j,i}$, $j = 1, \ldots, 4$ rows and $i = 1, \ldots, 4$ columns. For comparison also the array achievable subspace elements of the analytic model based system matrix $\dot{\mathbf{M}}_{N_A,N_C}$ from Ch. 2.2.1 are shown. High similarity between model and measurement can be observed for the matrix elements of the main diagonal up to approximately 5 kHz. Further it can be seen that the elements of the main diagonal $\dot{\mathrm{H}}_{1,1}, \dot{\mathrm{H}}_{2,2}, \dot{\mathrm{H}}_{3,3}, \dot{\mathrm{H}}_{4,4}$ are 20 dB above the erroneous elements in amplitude for most frequencies in the relevant range up to 2 kHz, which indicates a relatively good separation between the spherical harmonics 1 to 4. Nevertheless distortion is induced, that leads to deviations of the directivity patterns of the B-Format signals from the ideal omni-directional and figure of eight shaped patterns.

**First-Order Correction by System Inversion:** A way to overcome this problem and to re-establish the diagonal shape in the first 4 columns of the system matrix is to multiply Eq. 4.16 from the left side with an inverse $\dot{\boldsymbol{R}}_{N_A,N_A}$ of the system matrix $\dot{\mathbf{H}}_{N_A,N_C}$, which yields [Zau12]

$$\tilde{\boldsymbol{\chi}}_{N_A} = \dot{\boldsymbol{R}}_{N_A,N_A} \underbrace{\dot{\mathbf{H}}_{N_A,N_C}\ \boldsymbol{\phi}_{N_C}}_{\boldsymbol{\chi}_{N_A}}, \tag{4.17}$$




(a) W @ 350 Hz (b) Z @ 350 Hz (c) X @ 350 Hz

(d) W @ 2000 Hz (e) Z @ 2000 Hz (f) X @ 2000 Hz

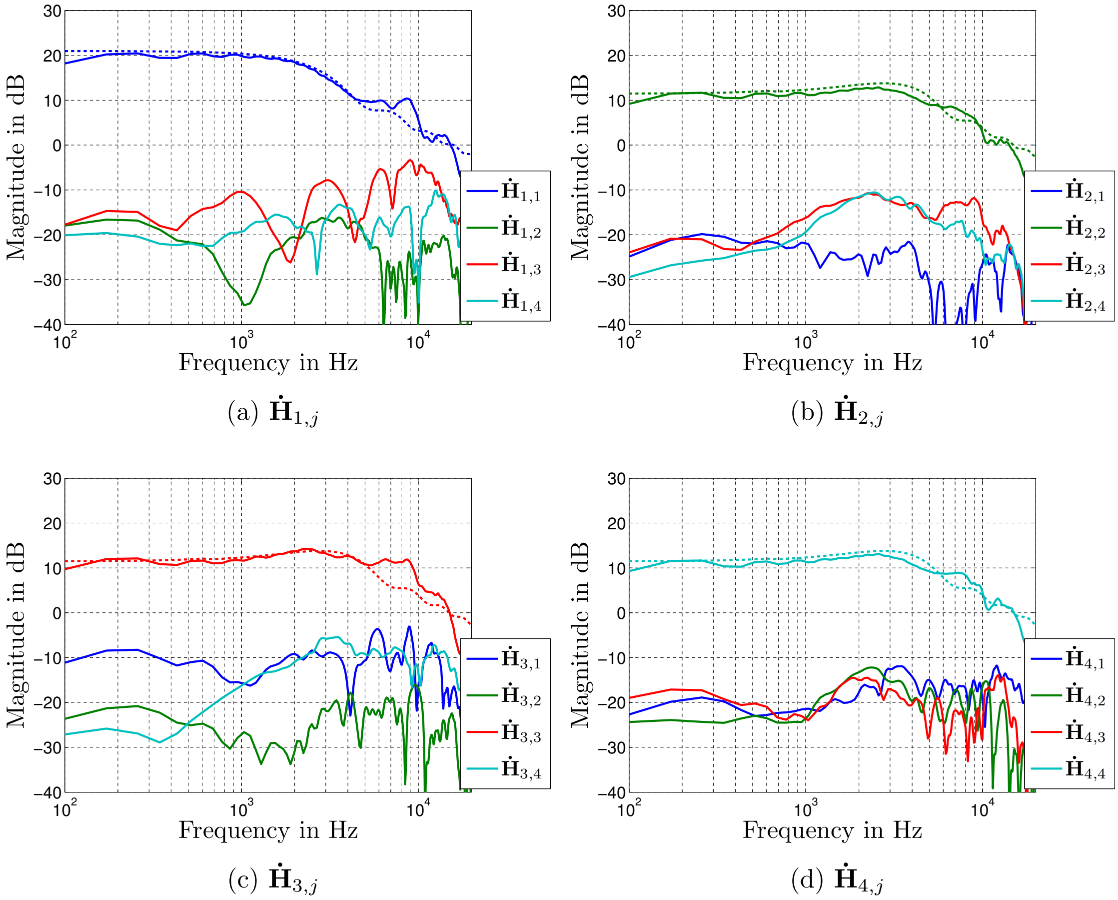
Figure 4.16: Directivity patterns of the measurement based B-Format components Z and X of $\boldsymbol{\chi}_{N_A}$ at different frequencies. (• → microphone positions). (Array 1)

where $\tilde{\boldsymbol{\chi}}_{N_A}$ denotes the filtered B-format signals. Since $\dot{\boldsymbol{R}}_{N_A,N_A}$ is a $(N_A+1)^2 \times (N_A+1)^2$ square matrix, the direct inverse can be used instead of the pseudo-inverse. If the excited spherical harmonics order is band-limited to order $N_A$, then the inversion is exact and $\tilde{\boldsymbol{\chi}}_{N_A} = \mathbf{I}\ \boldsymbol{\phi}_{N_A}$, where $\mathbf{I}$ is the identity matrix. The matrix $\dot{\boldsymbol{R}}_{N_A,N_A}$ is defined as

$$\dot{\boldsymbol{R}}_{N_A,N_A} = \left(\mathbf{Y}_{N_A}^{+}\mathbf{H}\mathbf{D}_{N_A}^{+}\right)^{-1}. \tag{4.18}$$

Fig. 4.18 shows the filtered system matrix $\dot{\mathbf{R}}_{N_A,N_A}\dot{\mathbf{H}}_{N_A,N_C}$ at a frequency of 2 kHz. As expected the array achievable subspace now possesses the shape of a diagonal matrix. The filters $\dot{\mathbf{R}}_{N_A,N_A}$ also change the higher order spherical harmonics resp. the spatial aliasing. Whether or not this is beneficial for the DOA estimation will be seen later on.

With all that said we want to evaluate the DOA estimation error of the Oktava 4D Ambient tetrahedral array. Utilizing the measured IRs $h_j(\boldsymbol{\theta}_l, f)$, with $j = 1, \ldots, 4$




(a) $\dot{\mathbf{H}}_{1,j}$ (b) $\dot{\mathbf{H}}_{2,j}$

(c) $\dot{\mathbf{H}}_{3,j}$ (d) $\dot{\mathbf{H}}_{4,j}$

Figure 4.17: $j = 1, \ldots, 4$ rows and first $i = 1, \ldots, 4$ columns of measurement based system matrix $\dot{\mathbf{H}}_{j,i}$ (solid lines) and model based system matrix $\dot{\mathbf{M}}_{j,i}$ (dotted lines) over frequency. (Array 1)

and $l = 1, \ldots, L$ that correspond to the $L = 576$ loudspeaker positions, the A- to B-Format conversion yields

$$\boldsymbol{\chi}_{N_A}(\boldsymbol{\theta}_l) = \mathbf{Y}_{N_A}^{+}\big(h_1(\boldsymbol{\theta}_l), h_2(\boldsymbol{\theta}_l), h_3(\boldsymbol{\theta}_l), h_4(\boldsymbol{\theta}_l)\big)^{\mathrm{T}}, \tag{4.19}$$

which leads to a total number of $L = 36 \times 16 = 576$ 4-element B-Format vectors $\boldsymbol{\chi}_{N_A}$ for the observed frequency $f$. For the sake of better readability the frequency $f$ is omitted in Eq. 4.19 and from now on. Intensity vector based DOA estimation is then performed with Eq. 2.1 for each B-Format vector $\boldsymbol{\chi}_{N_A}(\boldsymbol{\theta}_l) = [\mathrm{W}(\boldsymbol{\theta}_l), \mathrm{Y}(\boldsymbol{\theta}_l), \mathrm{Z}(\boldsymbol{\theta}_l), \mathrm{X}(\boldsymbol{\theta}_l)]^{\mathrm{T}}$, which yields $L$ DOA estimates written in spherical coordinates as $\hat{\boldsymbol{\theta}}_l \equiv (\hat{\varphi}_l, \hat{\vartheta}_l)$. The DOA estimation error in azimuth and zenith is then obtained by

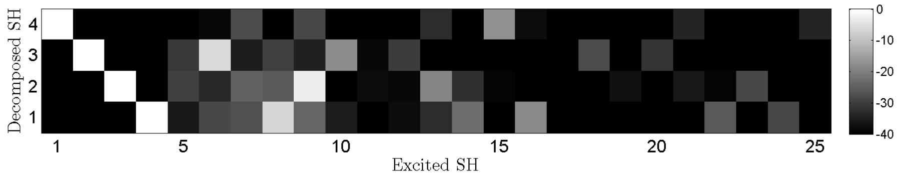



Figure 4.18: First 25 columns of filtered system matrix $\dot{\mathbf{R}}_{N_A,N_A}\dot{\mathbf{H}}_{N_A,N_C}$ at a frequency of 2000 Hz. (Array 1)

$$\Delta\varphi(\hat{\boldsymbol{\theta}}_l) = \varphi_l - \hat{\varphi}_l, \tag{4.20}$$

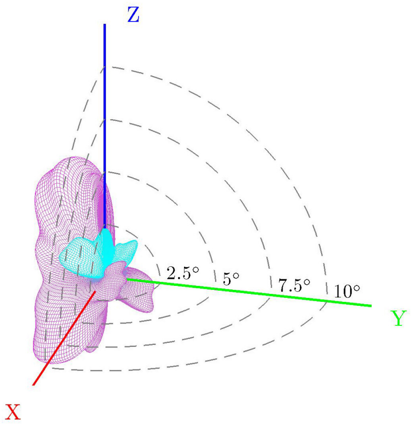
$$\Delta\vartheta(\hat{\boldsymbol{\theta}}_l) = \vartheta_l - \hat{\vartheta}_l, \tag{4.21}$$

where $\boldsymbol{\theta}_l \equiv (\varphi_l, \vartheta_l)$ corresponds to the true source direction of the $l^{th}$ loudspeaker. Fig. 4.19a,b shows the DOA estimation error in azimuth and zenith for a frequency of 350 Hz. Spherical interpolation is used for a finer resolution. At this low frequency the DOA estimation errors are mainly caused by the erroneous elements in the first 4 columns of the system matrix and not by spatial aliasing. Since azimuth errors around the equator are length preserving other than those around the poles, the weighting function

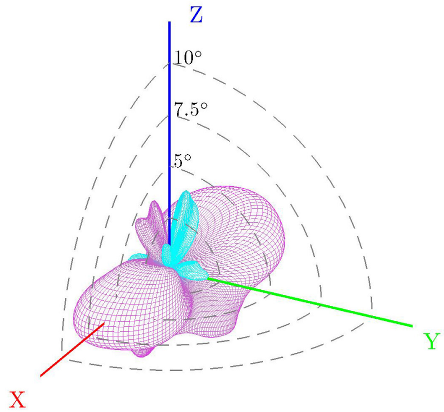
$$\mathcal{W}(\vartheta_l) = \sin(\vartheta_l), \tag{4.22}$$

is applied to azimuth errors of all DOA estimation error directivity plots in this work, where $\vartheta_l$ denotes the zenith angle of the $l^{th}$, $l = 1, \ldots, L$ measured source position.

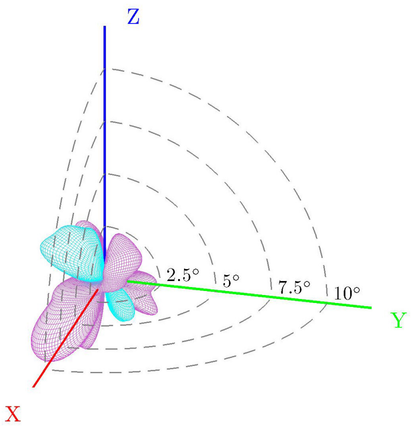
Applying the filters $\dot{\boldsymbol{R}}_{N_A,N_A}$ to the B-format signals $\boldsymbol{\chi}_{N_A}$ from Eq. 4.19 and using the filtered B-format signals $\tilde{\boldsymbol{\chi}}_{N_A} = \dot{\boldsymbol{R}}_{N_A,N_A}\ \boldsymbol{\chi}_{N_A}$ for the DOA estimation yields the DOA estimation errors shown in Fig. 4.19c,d. The filters $\dot{\boldsymbol{R}}_{N_A,N_A}$ completely remove the DOA estimation errors caused by erroneous elements in the first 4 columns of the system matrix. The remaining DOA estimation errors in Fig. 4.19c,d are caused by spatial aliasing of small but existent spherical harmonics of higher order, which are also filtered by $\dot{\boldsymbol{R}}_{N_A,N_A}$ and aliased into the decomposed spherical harmonics of the zeroth and first order.
The effect of the re-diagonalization filters on spatial aliasing can be seen in Fig. 4.20c,d, which shows the errors in azimuth and zenith for DOA estimation with Eq. 2.2 using B-format signals $\boldsymbol{\chi}_{N_A}(kR)$ and the filtered B-format signals $\tilde{\boldsymbol{\chi}}_{N_A}(kR)$ for a frequency



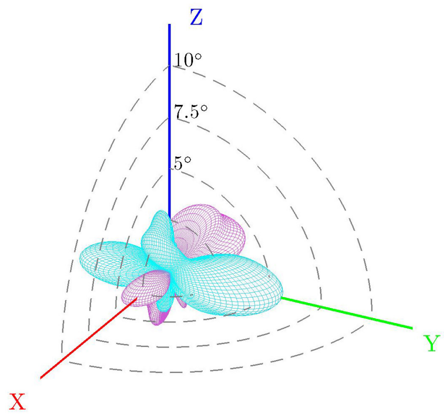

(a) $\Delta\varphi$, using $\boldsymbol{\chi}_{N_A}(kR)$




(b) $\Delta\vartheta$, using $\boldsymbol{\chi}_{N_A}(kR)$




(c) $\Delta\varphi$, using $\tilde{\boldsymbol{\chi}}_{N_A}(kR)$

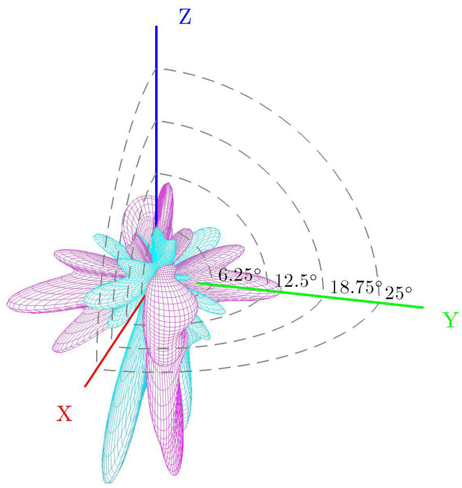



(d) $\Delta\vartheta$, using $\tilde{\boldsymbol{\chi}}_{N_A}(kR)$

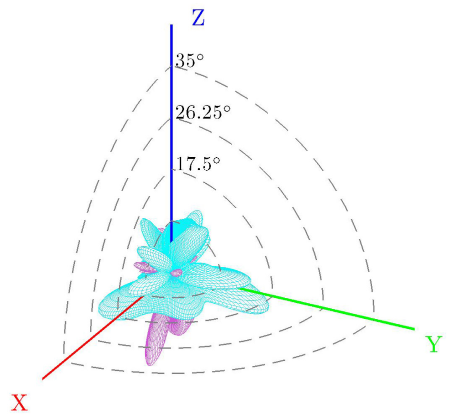
Figure 4.19: Error pattern of DOA estimation azimuth ($\Delta\varphi$) and zenith error ($\Delta\vartheta$) in °. DOA estimation is done according to Eq. 4.17; (a),(b) based on B-Format signals $\boldsymbol{\chi}_{N_A}$ and (c),(d) based on filtered B-Format signals $\tilde{\boldsymbol{\chi}}_{N_A}$ for $f = 350$ Hz and excitation order $N_C = 10$ (magenta: $\Delta\varphi, \Delta\vartheta < 0$. cyan: $\Delta\varphi, \Delta\vartheta \geq 0$). (Array 1)

of $f = 2000$ Hz. While the filters $\dot{\boldsymbol{R}}_{N_A,N_A}$ are beneficial at low frequencies, they increase the DOA estimation errors in azimuth and zenith for higher frequencies.

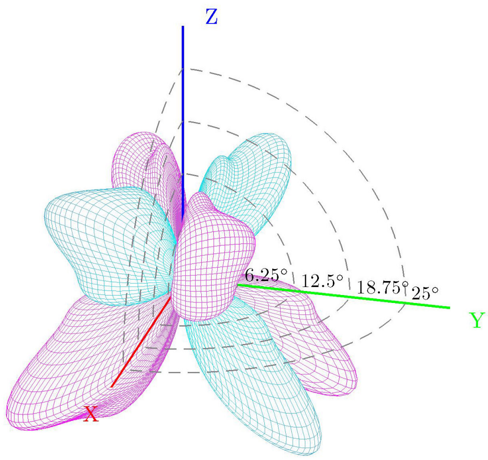
To get more overview of the effect of the radial filters on the DOA estimation we refer to the *weighted Mean Absolute Error* (wMAE) in azimuth and zenith, which



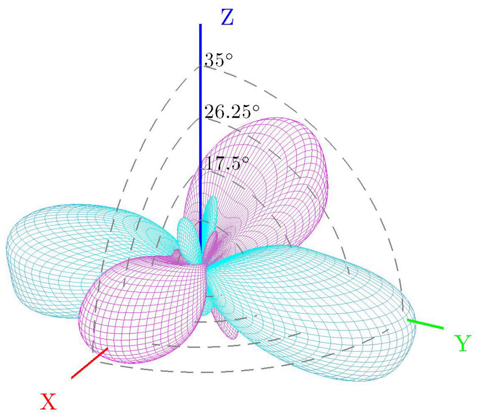

(a) $\Delta\varphi$, using $\boldsymbol{\chi}_{N_A}(kR)$




(b) $\Delta\vartheta$, using $\boldsymbol{\chi}_{N_A}(kR)$




(c) $\Delta\varphi$, using $\tilde{\boldsymbol{\chi}}_{N_A}(kR)$



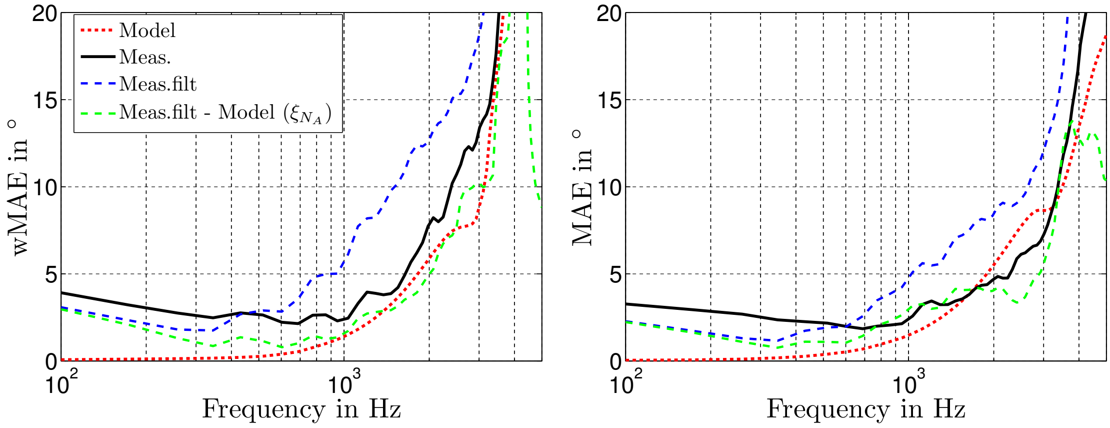

(d) $\Delta\vartheta$, using $\tilde{\boldsymbol{\chi}}_{N_A}(kR)$

Figure 4.20: Error pattern of DOA estimation azimuth ($\Delta\varphi$) and zenith error ($\Delta\vartheta$) in °. DOA estimation is done according to Eq. 4.17; (a),(b) based on B-Format signals $\boldsymbol{\chi}_{N_A}$ and (c),(d) based on filtered B-Format signals $\tilde{\boldsymbol{\chi}}_{N_A}$ for $f = 2000$ Hz and excitation order $N_C = 10$ (magenta: $\Delta\varphi, \Delta\vartheta < 0$. cyan: $\Delta\varphi, \Delta\vartheta \geq 0$). (Array 1)

consolidates the error information and is defined as

$$\Delta\varphi_{\text{wMAE}} = \frac{\sum_{l=1}^{L} |\mathcal{W}(\vartheta_l)\ \Delta\varphi_l|}{\sum_{l=1}^{L} |\mathcal{W}(\vartheta_l)|}, \tag{4.23}$$

where $\mathcal{W}(\vartheta_l)$ is the weight from Eq. 4.22, $l = 1, \ldots, L$ is the source position index and $\Delta\varphi_l$, $\Delta\vartheta_l$ denotes the azimuth and zenith DOA estimation error corresponding to the $l^{th}$ source position.




(a) $\Delta\varphi_{\mathrm{wMAE}}$

(b) $\Delta\vartheta_{\mathrm{MAE}}$

Figure 4.21: Weighted MAE in azimuth $\Delta\varphi_{\mathrm{wMAE}}$ and MAE in zenith $\Delta\vartheta_{\mathrm{MAE}}$ DOA estimation error over frequency based on different B-Format signals: Model $\boldsymbol{\chi}_{N_A}$ (Eq. 2.14), Meas.nofilt $\boldsymbol{\chi}_{N_A}$ (Eq. 4.16), Meas.filt $\tilde{\boldsymbol{\chi}}_{N_A}$ (Eq. 4.17), Meas.filt $\hat{\boldsymbol{\phi}}_{N_A}$ (Eq. 4.17) - Model $\boldsymbol{\xi}_{N_A}$ (Eq. 5.1). (Array 1)

Fig. 4.21 shows the azimuth and zenith wMAE from Eq. 4.23. The different curves correspond to different underlying B-format signals that have been used for the intensity vector based DOA estimation procedure from Eq. 2.1, where Model (red dotted curve) corresponds to the B-format signals $\boldsymbol{\chi}_{N_A}$ of the mathematical model from Eq. 2.14, Meas.nofilt (black solid curve) to the B-format signals $\boldsymbol{\chi}_{N_A}$ of the measurement from Eq. 4.16 and Meas.filt (blue dashed line) to the filtered B-format signals $\tilde{\boldsymbol{\chi}}_{N_A}$ of the measurement. The green dashed curve is explained in Ch. 5.
The figure shows that the re-diagonalization filters lead to an error reduction up to approximately 500 Hz and then cause increased errors for higher frequencies, which can be attributed to the fact that the filters modify the spatial aliasing, which is high for high frequencies, in a way that negatively impacts the DOA estimation. Even with applied re-diagonalization filters, a mean azimuth and zenith error of 2° to 3° remains,[12] which can lead to significant errors in the 3-dimensional localization procedure and therefore a more rigorous DOA error correction is described in Ch. 6. Compared with the filtered zenith wMAE curve, the non-filtered version is spectrally

[12]E.g. an azimuth error of 2° to 3° can lead to a 3-dimensional localization error of 30 cm assuming a scenario of 2 microphone arrays with bearing lines that intersect with an angle of 30°.

flat over a wider frequency range, which is desirable for our purpose of developing a broadband localization algorithm, since it makes further correction in Ch. 6 easier. Thus the re-diagonalization filters were not implemented, but might be interesting for a low-pass localization procedure.

# Chapter 5

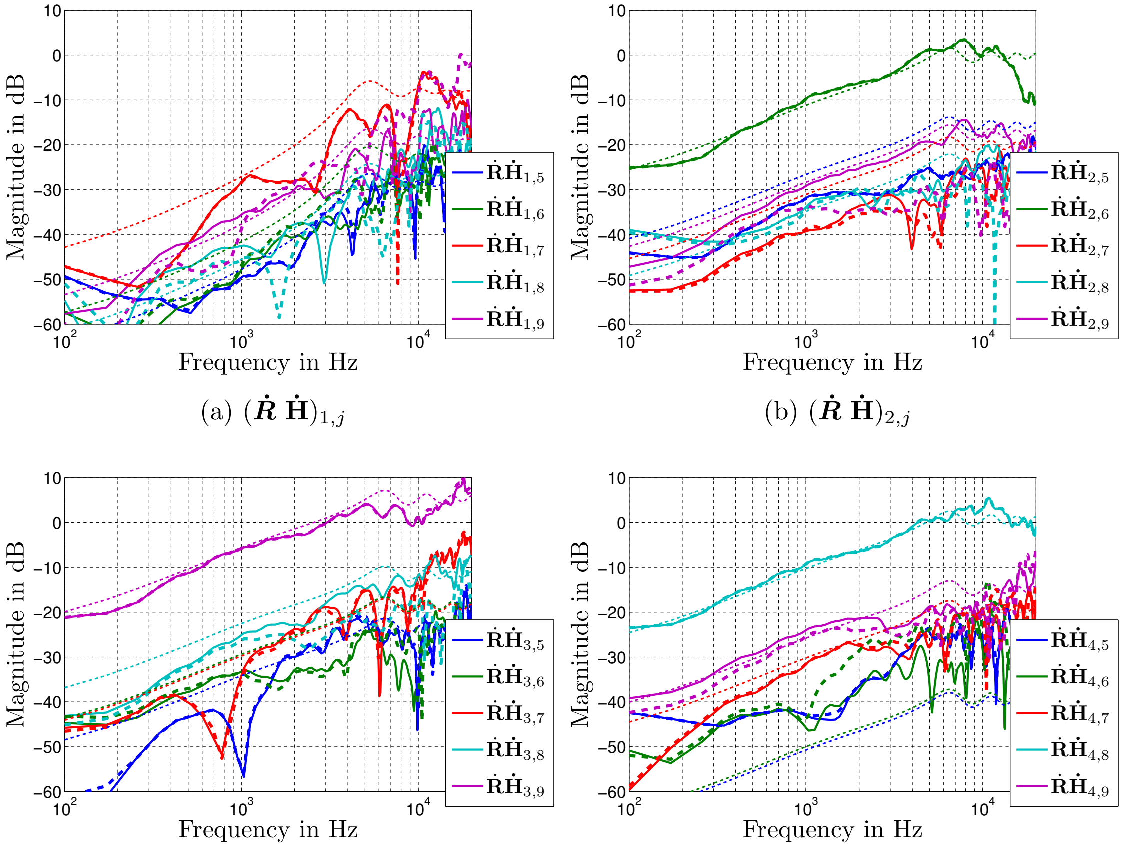
# Comparison of Model and Measurement

In this chapter we investigate whether the model from Ch. 2.2.1 is able to reproduce the DOA estimation errors of the measurement from Fig. 4.20. This would be convenient as it allows partial error correction to users that do not have the equipment or time to conduct sophisticated multi-source IR measurements as in Ch. 4.

From Fig. 2.7 we know that the array achievable subspace of the system matrix $\dot{\mathbf{M}}_{N_A,N_C}$ is a diagonal matrix, since the model is based on the assumptions of ideal microphones, i.e. the model is not able to reproduce the errors caused by erroneous elements in the array achievable subspace of the measurement (see Fig. 4.15). We therefore want to investigate whether the DOA estimation errors caused by spatial aliasing can be reproduced correctly.
According to Fig. 2.7 impinging spherical harmonics of second order cause the main part of the total spatial aliasing. The system matrix based on measurements in Fig. 4.15 shows a similar result, where the matrix elements corresponding to impinging spherical harmonics $6, 8$ and $9$ exhibit the highest magnitude apart from the elements of the main diagonal.[1]

To gain more insight into spatial aliasing, Fig. 5.1 shows the magnitude responses of all elements of columns $5 \ldots 9$ of the system matrix $\dot{\mathbf{H}}_{N_A,N_C}$, i.e. the spatial aliasing of the spherical harmonics of $2^{nd}$ order onto the decomposed spherical harmonics $1 \ldots 4$ on the microphone side. Since we want to compare the spatial aliasing in each row to the magnitude of the corresponding main diagonal element, e.g. the frequency

[1]For high frequencies the matrix elements of second order even exceed the elements of the main diagonal in magnitude as can be seen in Fig. 5.1.




(a) $(\dot{\boldsymbol{R}}\,\dot{\mathbf{H}})_{1,j}$ (b) $(\dot{\boldsymbol{R}}\,\dot{\mathbf{H}})_{2,j}$

(c) $(\dot{\boldsymbol{R}}\,\dot{\mathbf{H}})_{3,j}$ (d) $(\dot{\boldsymbol{R}}\,\dot{\mathbf{H}})_{4,j}$

Figure 5.1: Rows $j = 1, \ldots, 4$ and columns $i = 5, \ldots, 9$ of system matrix $\dot{\mathbf{H}}_{N_A,N_C}$ (solid curves), filtered system matrix $(\dot{\boldsymbol{R}}\,\dot{\mathbf{H}})_{j,i}$ (dashed curves) and analytic model (dotted curves) over frequency. Each row is normalized to its main diagonal element.

response of system matrix element $H_{3,9}$ with respect to main diagonal element $H_{3,3}$, the magnitude responses of each row are normalized to the magnitude responses of the corresponding main diagonal elements. For comparison the magnitude responses of columns $5 \ldots 9$ of the analytical model based system matrix $\dot{\mathbf{M}}_{N_A,N_C}$ are shown as dotted lines. It can be verified, that the dominating spherical harmonic of $2^{nd}$ order in each row is the same for model $\dot{\mathbf{M}}_{N_A,N_C}$ and measurement $\dot{\mathbf{H}}_{N_A,N_C}$, and the corresponding frequency responses exhibit a highly similar tendency over the whole frequency spectrum, which indicates that the analytical model is able to qualitatively describe the spatial aliasing of the measurement.

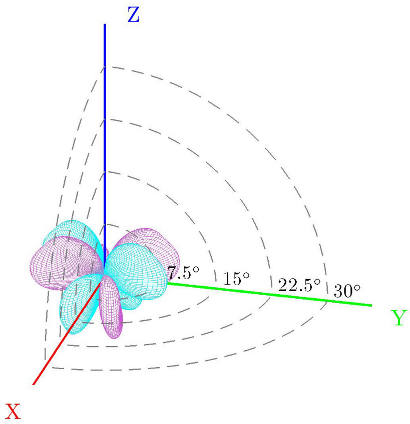
To generate a model-based DOA estimation error that is comparable to Fig. 4.19 and Fig. 4.20, we also use point source excitation for the analytic model. This is done by inserting the wave spectrum $\boldsymbol{b}_{N_C} = -ik\,\mathrm{diag}\{\boldsymbol{h}_{N_C}(kr_0)\}\mathbf{y}_{N_C}(\boldsymbol{\theta}_i)$ of a point source from direction $\boldsymbol{\theta}_i$ with a distance of $r_0 = 1.4$ m to the microphone array into Eq. 2.13 and then using the resulting spherical wave spectrum $\boldsymbol{\chi}_{N_A}(kR)$ for the DOA estimation, where $\boldsymbol{h}_{N_C}(kr_0) = [h_0,\ h_1,\ h_1,\ h_1,\ h_2, \ldots,\ h_{N_C}]^T$ is a vector containing the Hankel functions $h_n$ of the second kind up to order $n = N_C$. The DOA estimation error of the analytic model based on the resulting spherical wave spectrum $\boldsymbol{\chi}_{N_A}(kR)$ is illustrated in Fig. 5.2a,b. Comparing Fig. 5.2a,b and Fig. 4.20c,d we can observe that the error patterns of model and measurement differ significantly. The assumption proved to be invalid that the DOA estimation errors caused by spatial aliasing dominate in the system matrix $\dot{\mathbf{H}}_{N_A,N_C}$ at high frequencies. Hence reproduction of the measured DOA errors via the model is not valid. A superior prediction of the DOA error patterns of model and measurement is obtained by multiplying the term $\mathbf{P}_{N_A}^{-1}\mathrm{diag}\{\frac{1}{\boldsymbol{\rho}_{N_A}(kR)}\}$ from the left to the spherical wave spectrum from Eq. 2.14, which yields the *plane wave decomposition* as [Raf04]

$$\boldsymbol{\xi}_{N_A}(kR) = \mathbf{P}_{N_A}^{-1}\mathrm{diag}\{\frac{1}{\boldsymbol{\rho}_{N_A}(kR)}\}\underbrace{\mathbf{Y}_{N_A}^{+}\mathbf{Y}_{N_C}\mathrm{diag}\{\boldsymbol{\rho}_{N_C}(kR)\}\boldsymbol{b}_{N_C}}_{\boldsymbol{\chi}_{N_A}}, \tag{5.1}$$

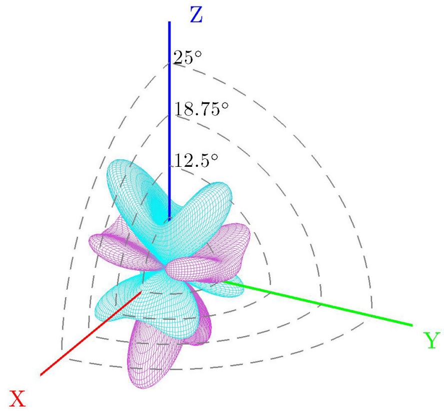
with

$$\mathbf{P}_{N_A} = \begin{pmatrix} 4\pi & 0 & \cdots & 0 \\ 0 & 4\pi\mathrm{i} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & 4\pi\mathrm{i}^{N_A}, \end{pmatrix}, \tag{5.2}$$

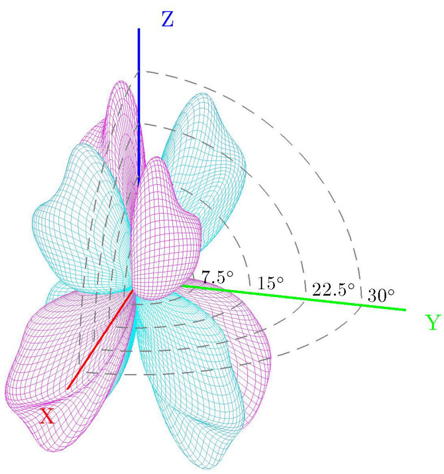
where $\boldsymbol{\rho}_{N_C}(kR)$ denotes a vector containing the radial functions from Eq. 2.11 up to order $n = N_C$. DOA estimation errors based on $\boldsymbol{\xi}_{N_A}(kR)$ are illustrated in Fig. 5.2c,d. The wMAE of the difference between the DOA estimation errors based on the filtered measurement data $\tilde{\boldsymbol{\chi}}_{N_A}(kR)$ and on plane wave decomposition model data $\boldsymbol{\xi}_{N_A}(kR)$ is illustrated in azimuth and zenith in Fig. 4.21. Compared to the wMAE curve of the filtered measurement data, it can be observed that for frequencies higher than approximately 300 Hz a huge part of the DOA estimation errors caused by spatial aliasing can be removed, i.e. can be reproduced correctly by the analytic model. However the microphone array dependent re-diagonalization filters $\dot{\boldsymbol{R}}$ are needed to obtain $\tilde{\boldsymbol{\chi}}_{N_A}(kR)$. DOA correction based on the model data is not implemented in this work, instead the approach in Ch. 6 is used.



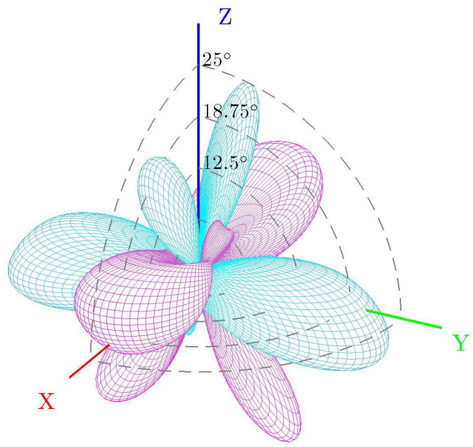

(a) $\Delta\varphi$, using $\boldsymbol{\chi}_{N_A}(kR)$




(b) $\Delta\vartheta$, using $\boldsymbol{\chi}_{N_A}(kR)$




(c) $\Delta\varphi$, using $\boldsymbol{\xi}_{N_A}(kR)$




(d) $\Delta\vartheta$, using $\boldsymbol{\xi}_{N_A}(kR)$

Figure 5.2: Error pattern of DOA estimation azimuth ($\Delta\varphi$) and zenith error ($\Delta\vartheta$) in °. DOA estimation is done according to Eq. 4.17; (a),(b) based on B-Format signals $\boldsymbol{\chi}_{N_A}$ (model Eq. 2.14) and (c),(d) based on filtered B-Format signals $\tilde{\boldsymbol{\chi}}_{N_A}$ (eq: doa est. single: compare mod meas: pl wave decomp) for $f = 2000$ Hz and excitation order $N_C = 10$ (magenta: $\Delta\varphi, \Delta\vartheta < 0$. cyan: $\Delta\varphi, \Delta\vartheta \geq 0$).

# Chapter 6

# DOA Correction

The filters $\dot{\boldsymbol{R}}_{N_A,N_A}$ introduced in Ch. 4.5.3 manage to improve the decomposition of the soundfield into spherical harmonics and subsequently the accuracy of the DOA estimations for low frequencies, where the spherical harmonics of higher orders $n > N_A$ are small in magnitude and therefore little spatial aliasing occurs. At higher frequencies, where spatial aliasing increases, the filters worsen the DOA estimation performance. Instead a different approach is used to minimize the DOA estimation errors, where DOA estimation is performed for $L$ different source directions and the DOA estimation errors are used to generate two correction matrices (for each frequency $f_\mathrm{k}$), one for the azimuth- and one for the zenith angle. The approach is presented in this chapter.

In Ch. 4.5.3 the measured IRs of $L = 576$ loudspeaker positions are used for A- to B-Format conversion (Eq. 4.19) and with the obtained B-Format signals DOA estimation is performed, which yields the DOA estimates in spherical coordinates $\hat{\boldsymbol{\theta}}_l \equiv (\hat{\varphi}_l, \hat{\vartheta}_l)$, $l = 1, \ldots, L$ (loudspeaker index) at a single frequency. Subtracting the $l^{th}$ DOA estimate from the true loudspeaker direction $\boldsymbol{\theta}_l \equiv (\varphi_l, \vartheta_l)$ (Eq. 4.20) yields the DOA error in azimuth $\Delta\varphi(\hat{\boldsymbol{\theta}}_l)$ and zenith $\Delta\vartheta(\hat{\boldsymbol{\theta}}_l)$, which can be mapped to the corresponding DOA estimation $\hat{\boldsymbol{\theta}}_l$ and the corrected source direction are obtained through $\varphi_l = \hat{\varphi}_l + \Delta\varphi(\hat{\boldsymbol{\theta}}_l)$ and $\vartheta_l = \hat{\vartheta}_l + \Delta\vartheta(\hat{\boldsymbol{\theta}}_l)$. Since we want to estimate an error correction to an arbitrary DOA estimation $\hat{\boldsymbol{\theta}}$ that might lie in between the $L = 576$ measured DOA estimations $\hat{\boldsymbol{\theta}}_l$, interpolation is used.
The functions that we want to interpolate are $\Delta\varphi(\hat{\boldsymbol{\theta}}_l)$ and $\Delta\vartheta(\hat{\boldsymbol{\theta}}_l)$ at each frequency $f_\mathrm{k}$. In contrast to the loudspeaker directions $\boldsymbol{\theta}_l$, which are distributed uniformly on the sphere according to an equi-angular sampling scheme (see fig. 6.1, black dots), the estimated directions $\hat{\boldsymbol{\theta}}_l$ are distributed irregularly because of the estimation errors. Therefore we have a set of data $\Delta\varphi(\hat{\boldsymbol{\theta}}_l)$ and $\Delta\vartheta(\hat{\boldsymbol{\theta}}_l)$ sampled at scattered directions $\hat{\boldsymbol{\theta}}_l$.

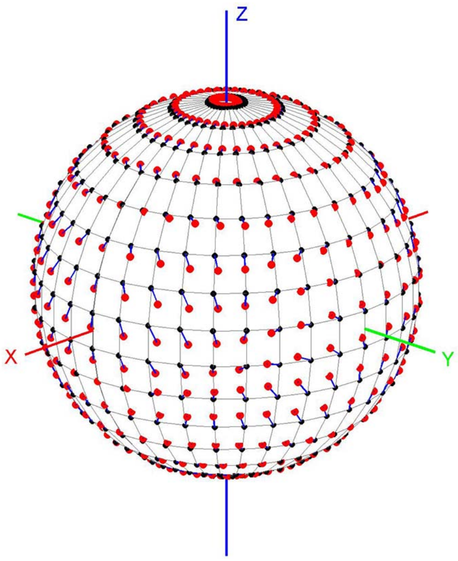
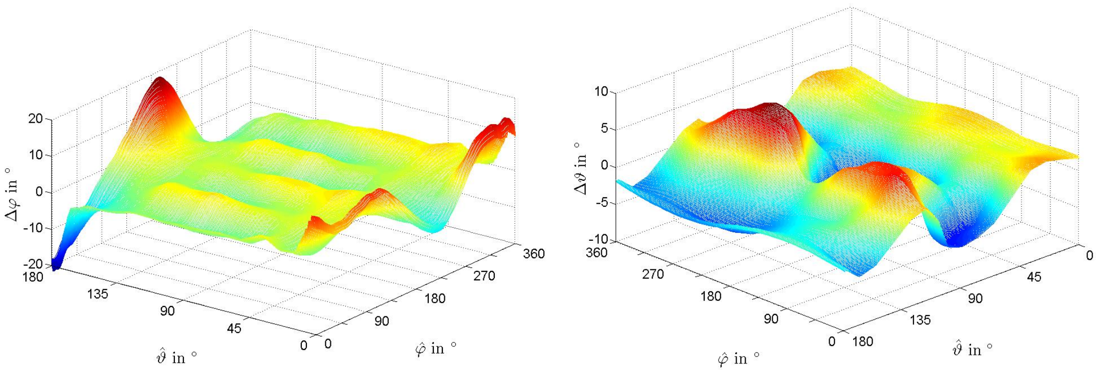



Figure 6.1: Estimated DOA $\hat{\boldsymbol{\theta}}_l$ (●), real DOA $\boldsymbol{\theta}_l$ (●) and position Offset (—) at a frequency of $f = 1$ kHz.

For interpolation we use Matlab's *Triscatteredinterp* function that creates an interpolant, which fits a surface defined by values of scattered data points. Interpolation is done separately for the two functions $\Delta\varphi(\hat{\boldsymbol{\theta}}_l)$ and $\Delta\vartheta(\hat{\boldsymbol{\theta}}_l)$. Due to the fact that the functions change with frequency, we generate interpolations $\Delta\varphi(\hat{\boldsymbol{\theta}})$ and $\Delta\vartheta(\hat{\boldsymbol{\theta}})$ for multiple frequencies $f_\mathrm{k}$, where k denotes the frequency index $\mathrm{k} = 1, \ldots, \mathrm{K}$. The total number K of observed frequencies corresponds to the DFT length that was used in DOA estimation (see Ch. 2.1), i.e. DOA estimations of frequency index k can be corrected with the corresponding two interpolated functions $\Delta\varphi_\mathrm{k}(\hat{\boldsymbol{\theta}}_\mathrm{k})$ and $\Delta\vartheta_\mathrm{k}(\hat{\boldsymbol{\theta}}_\mathrm{k})$, which will be denoted as *correction matrices*.

Assuming a DFT length of $\mathrm{K} = 4096$, storing the interpolated functions can be memory consuming, therefore $\Delta\varphi_\mathrm{k}(\hat{\boldsymbol{\theta}}_\mathrm{k})$ and $\Delta\vartheta_\mathrm{k}(\hat{\boldsymbol{\theta}}_\mathrm{k})$ are first evaluated at a dense uniform grid of $360 \times 180$ ($\Delta\hat{\varphi} = \Delta\hat{\vartheta} = 1°$) and then transformed by a 2-dimensional Fourier transform. Sufficient accuracy is achieved if only the first 20 coefficients are kept for the $\hat{\varphi}$ direction and 80 coefficients for the $\hat{\vartheta}$ direction, which results in 100 coefficients that have to be stored for each frequency index k.[1] Fig. 6.2 shows the two angularly band-limited correction matrices $\Delta\varphi_\mathrm{k}(\hat{\boldsymbol{\theta}}_\mathrm{k})$ and $\Delta\vartheta_\mathrm{k}(\hat{\boldsymbol{\theta}}_\mathrm{k})$ for a frequency of $f_\mathrm{k} = 1$ kHz.

[1]Compared to the azimuth direction, the number of coefficients in the zenith direction has to be higher, since the spectrum is not band-limited here due to jumps between $\hat{\theta} = 0°$ and $\hat{\theta} = 180°$. A small number of coefficients would lead to a strong ripple (Gibbs phenomenon).




(a) $\Delta\varphi(\hat{\boldsymbol{\theta}})$ (b) $\Delta\vartheta(\hat{\boldsymbol{\theta}})$

Figure 6.2: Correction matrices $\Delta\varphi(\hat{\boldsymbol{\theta}})$,$\Delta\vartheta(\hat{\boldsymbol{\theta}})$ in °, with $\hat{\varphi} \in [0°, 360°)$ and $\hat{\vartheta} \in [0°, 180°)$ at a frequency of $f = 1$ kHz.

Fig. 6.3 shows the measured $L = 576$ DOA estimations $\hat{\boldsymbol{\theta}}_l$ from Fig. 6.1 after DOA correction with the two correction matrices. Matlab's *Triscatteredinterp* function works in a way that it interpolates $\Delta\varphi_{\mathrm{k}}(\hat{\boldsymbol{\theta}}_{\mathrm{k}})$ and $\Delta\vartheta_{\mathrm{k}}(\hat{\boldsymbol{\theta}}_{\mathrm{k}})$ in between the scattered data points. At the scattered data points the interpolation matches exactly the given data point values. The small remaining offsets in Fig. 6.3 are therefore caused by the band-limitation to 100 Fourier coefficients. An alternative to Matlab's *Triscattered-interp* algorithm is the use of spherical interpolation. In comparison the combination of *Triscatteredinterp* and bandlimitation in the 2-dimensional Fourier domain yields better results and is therefore used.

The presented method for DOA correction works as long as any DOA estimation $\hat{\boldsymbol{\theta}}_{\mathrm{k}}$ can be uniquely mapped to an estimation error $\Delta\varphi_{\mathrm{k}}(\hat{\boldsymbol{\theta}}_{\mathrm{k}})$ and $\Delta\vartheta_{\mathrm{k}}(\hat{\boldsymbol{\theta}}_{\mathrm{k}})$. This also means that, assuming acoustic sources on a grid as in Fig. 6.1 (black dots), the DOA estimations $\boldsymbol{\theta}_k$ of neighboring source directions will not overtake each other. A graphical way to show this is to unroll the DOA directions illustrated on the sphere onto 2 dimensions and to connect the DOA estimations through lines. The property of unique mapping mentioned before is violated if the lines intersect. The lines are illustrated in Fig. 6.4 for frequencies $f_{\mathrm{k}} = 1$ kHz and $f_{\mathrm{k}} = 2$ kHz. Whereas for $f_{\mathrm{k}} = 1$ kHz the lines are clearly separated, for $f_{\mathrm{k}} = 2$ kHz they do intersect in some areas. However the lines only intersect in moderation and $f_{\mathrm{k}} = 2$ kHz denotes the maximum frequency where interpolation with *Triscatteredinterp* and thus DOA correction is still possible. In the subsequent observations 2 kHz is assumed to be the upper frequency limit.

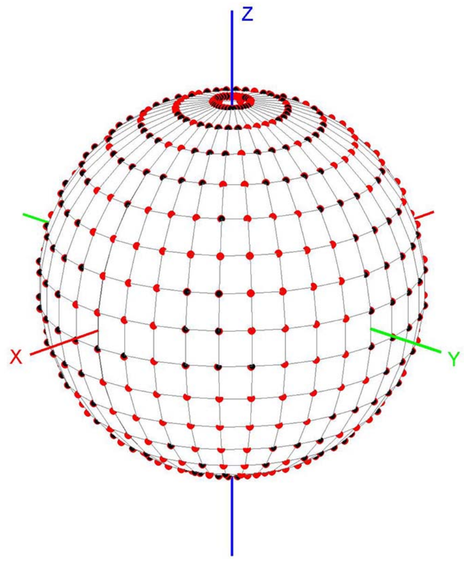
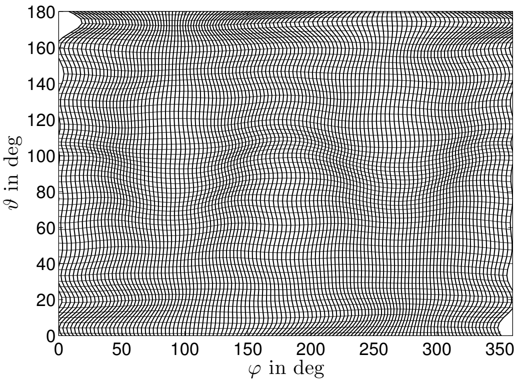
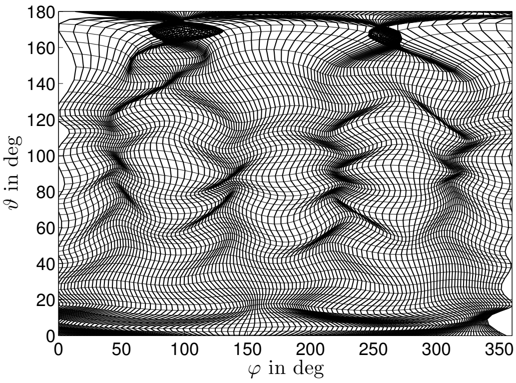



Figure 6.3: Estimated DOA $\hat{\boldsymbol{\theta}}_l$ after DOA correction (●), real DOA $\boldsymbol{\theta}_l$ (●) and position Offset (—) at a frequency of $f = 1$ kHz.




(a) $f = 1000\ Hz$

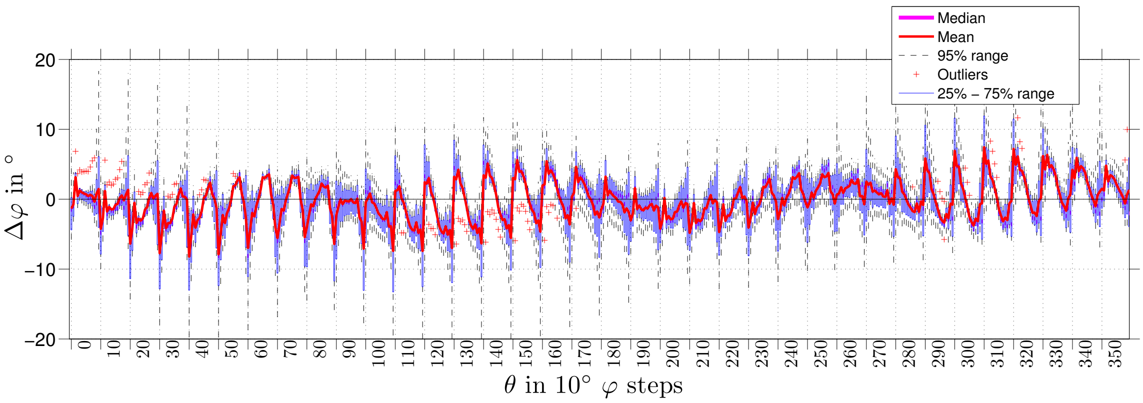



(b) $f = 2000\ Hz$

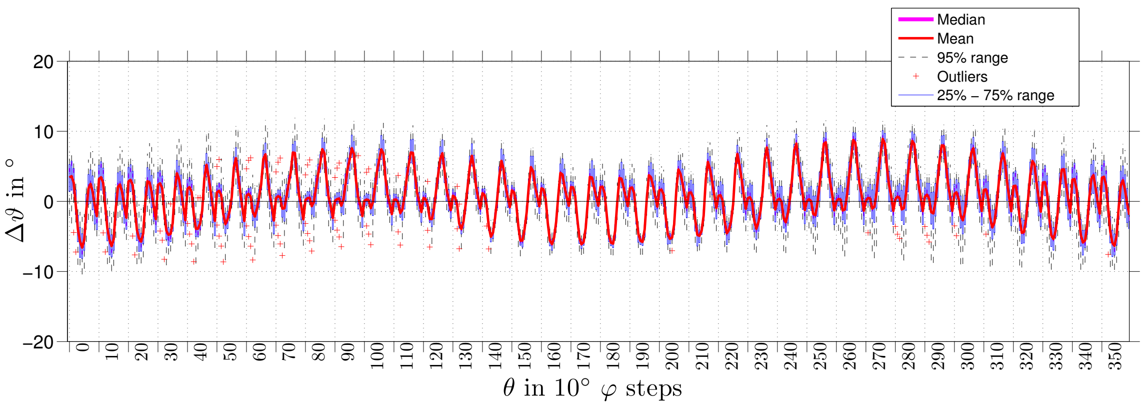
Figure 6.4: Localization lines for different frequencies $f$.

In terms of practical implementation, an important point in the DOA correction procedure is, whether the correction matrices of one microphone array can also be applied to another, as universal correction matrices would be handy especially if the equipment needed to measure the direction dependent IRs of a microphone array is not available. The DOA estimation errors differ significantly among the microphone arrays, which can be explained by the mismatch in the directivity patterns of the microphones. Fig. 6.5 shows the distribution of the DOA estimations errors for all $M = 7$ used microphone arrays at the source positions $\boldsymbol{\theta}$.

However by observing the mean and median value in in Fig. 6.5 a systematic estima-




(a) Distribution of $\Delta\varphi(\boldsymbol{\theta})$




(b) Distribution of $\Delta\vartheta(\boldsymbol{\theta})$

Figure 6.5: Distribution of $\Delta\varphi(\boldsymbol{\theta})$, $\Delta\vartheta(\boldsymbol{\theta})$ for all 7 microphone arrays at $f = 1$ kHz. A grid layout is used with $\varphi = 0, 10, \ldots, 350°$ and $\vartheta = 16.875, 28.125, \ldots, 163.125°$, which yields 36 meridians and 14 latitude circles. E.g. the values in between the ticks next to label 0 on the x axis describe the 14 latitude angles of loudspeakers 2 to 15 at the meridian $\varphi = 0°$.

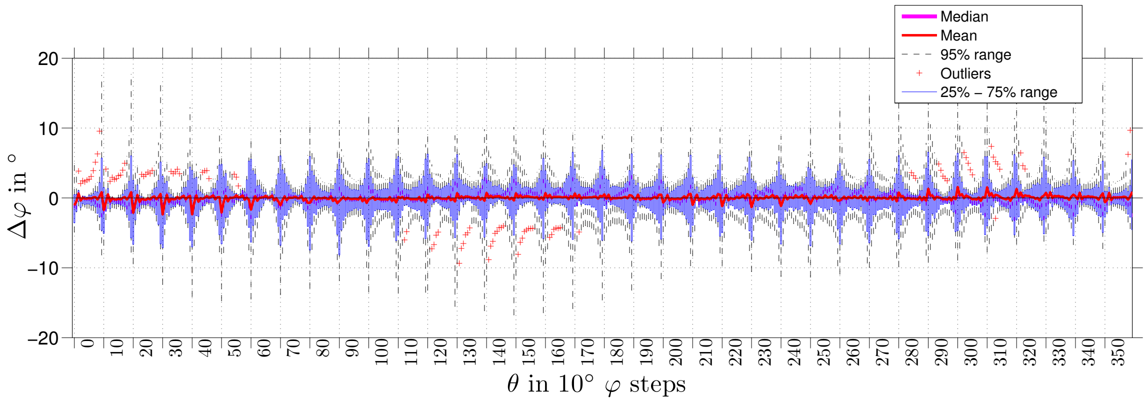
tion error can be seen that is contained in $\Delta\varphi(\boldsymbol{\theta})$ and $\Delta\vartheta(\boldsymbol{\theta})$ of all microphone arrays and the systematic error will also be contained in $\Delta\varphi(\hat{\boldsymbol{\theta}})$ and $\Delta\vartheta(\hat{\boldsymbol{\theta}})$.[2]
We make use of this and generate the mean estimation error matrices

$$\overline{\Delta\varphi}_{\mathrm{k}}(\hat{\boldsymbol{\theta}}_l) = \frac{1}{M}\sum_{m=1}^{M} \Delta\varphi_{\mathrm{k}}^{m}(\hat{\boldsymbol{\theta}}_l) \tag{6.1}$$

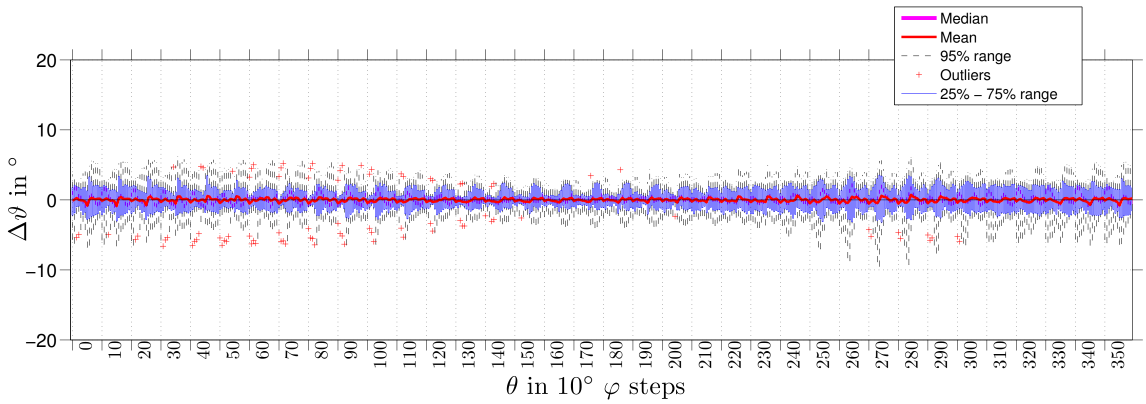
$$\overline{\Delta\vartheta}_{\mathrm{k}}(\hat{\boldsymbol{\theta}}_l) = \frac{1}{M}\sum_{m=1}^{M} \Delta\vartheta_{\mathrm{k}}^{m}(\hat{\boldsymbol{\theta}}_l), \tag{6.2}$$

[2]Note that observing the meridians ($\varphi = \{40, 130, 220, 310\}$) in Fig. 6.5a in between the microphone on-axis positions it can be seen that the mean value is drawn to the microphones.

where $m = 1, \ldots, M$ is the microphone array index. Analog to $\Delta\varphi_\mathrm{k}(\hat{\boldsymbol{\theta}}_l)$ and $\Delta\vartheta_\mathrm{k}(\hat{\boldsymbol{\theta}}_l)$, $\overline{\Delta\varphi}_\mathrm{k}(\hat{\boldsymbol{\theta}}_l)$ and $\overline{\Delta\vartheta}_\mathrm{k}(\hat{\boldsymbol{\theta}}_l)$ are interpolated and band-limited, which yields two mean correction matrices $\overline{\Delta\varphi}_\mathrm{k}(\hat{\boldsymbol{\theta}})$ and $\overline{\Delta\vartheta}_\mathrm{k}(\hat{\boldsymbol{\theta}})$ for each frequency index k. Using the matrices for DOA correction on the $36 \times 14$ grid points illustrated in Fig. 6.5 yields the the error variance in Fig. 6.6 that has been disposed of the systematic mean error to a great extend. However this does not affect the variance of the DOA estimation errors among the arrays, especially near the poles (see fig. 6.6a). These errors cannot be removed by a universal correction matrix as they are different for each microphone array, but can be removed completely with the individual (array dependent) correction matrices, which will be used for DOA correction in Ch. 7.




(a) Distribution of $\Delta\varphi(\boldsymbol{\theta})$




(b) Distribution of $\Delta\vartheta(\boldsymbol{\theta})$

Figure 6.6: Distribution of $\Delta\varphi(\boldsymbol{\theta})$, $\Delta\vartheta(\boldsymbol{\theta})$ for all 7 microphone arrays at $f = 1$ kHz after DOA error correction with the mean correction matrices $\overline{\Delta\varphi}(\hat{\boldsymbol{\theta}})$ and $\overline{\Delta\vartheta}(\hat{\boldsymbol{\theta}})$. A grid layout is used with $\varphi = 0, 10, \ldots, 350°$ and $\vartheta = 16.875, 28.125, \ldots, 163.125°$, which yields 36 meridians and 14 latitude circles. E.g. the values in between the ticks next to label 0 on the x axis describe the 14 latitude angles of loudspeakers 2 to 15 at the meridian $\varphi = 0°$.

# Chapter 7

# Experimental Evaluation

In this chapter the localization methods presented in the preceding chapters are evaluated through a number of measurements that were conducted with multiple spatially distributed Oktava 4D-Ambient tetrahedral microphone arrays. The measurements took place at the IEM CUBE, which was also used for the measurements described in Ch. 4. The CUBE is a middle sized multi-purpose room with dimensions of approximately 10 m ×12 m ×4 m, which is used as a lab and since it is equipped with a hemispherical array of 24 loudspeakers and an optical tracking system, it is often used as a venue for electro-acoustic performances and demonstrations [ZSR03, Fre10].

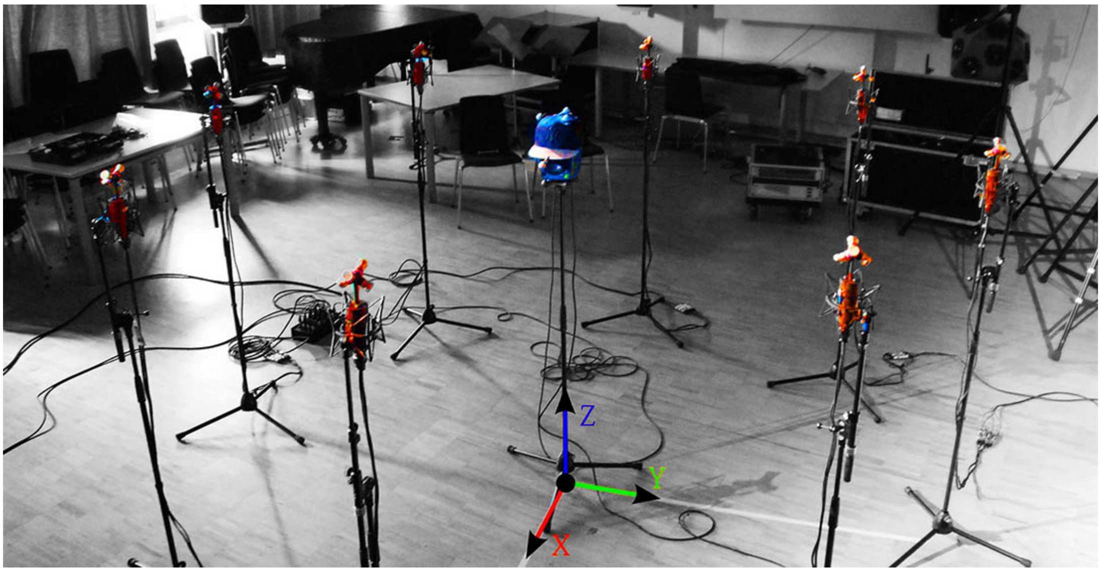
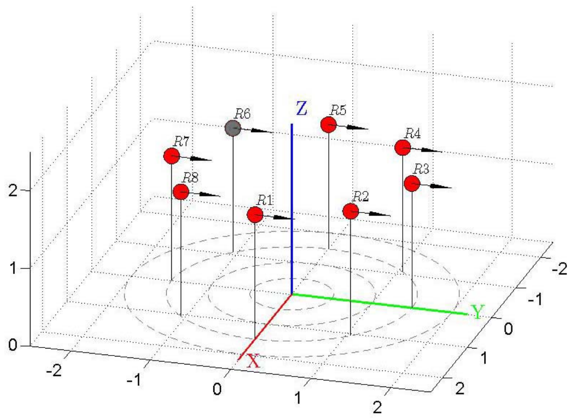



Figure 7.1: Microphone array setup and loudspeaker placed at source position $S_1$. A cap equipped with optical tracking sensors was placed on top of the loudspeaker to determine its position via the optical tracking system.

The optical tracking system was used for all positioning measurements, as it provides a highly accurate and convenient way of measuring static microphone array and source positions. Also positions of moving sources can be tracked easily. The hemispherical loudspeaker array was not used. Further the room provides sound absorbing curtains on the window side that were closed to reduce the reverberation time to some degree. The reverberation time with drawn curtains amounts to approximately $T_{60} \approx 0.65$ s.

## 7.1 Measurement Setup

To describe microphone and source positions, a 3-dimensional global right-handed Cartesian coordinate system was defined, with the origin located on the floor approximately in the middle of the room and the orientations of the coordinate axes are defined as illustrated in Fig. 7.1. The microphone array setup is illustrated in Fig. 7.2a with a total number of 8 tetrahedral microphone arrays. The microphone arrays are positioned on a circle with a radius of $r_0 = 1.5$ m in $\Delta\varphi = +45°$ azimuth steps starting from microphone array 1 at $\varphi = 0°$ on the x-axis at a height of $z = 1.6$ m. Due to excessive DOA estimation errors, microphone array $R_6$ is not implemented. The measured array positions are listed in Tab. 7.2b.




(a) Array positions (•) and array orientation (→)

| | $x$ [m] | $y$ [m] | $z$ [m] |
|---|---|---|---|
| $\mathbf{r}_1$ | 1.5 | 0 | 1.6 |
| $\mathbf{r}_2$ | 1.061 | 1.061 | 1.6 |
| $\mathbf{r}_3$ | 0 | 1.5 | 1.6 |
| $\mathbf{r}_4$ | -1.061 | 1.061 | 1.6 |
| $\mathbf{r}_5$ | -1.5 | 0 | 1.6 |
| $\mathbf{r}_7$ | 0 | -1.5 | 1.6 |
| $\mathbf{r}_8$ | 1.061 | -1.061 | 1.6 |

(b) Array positions in x,y,z coordinates

Figure 7.2: Microphone array setup

For all measurements, the microphone arrays were positioned such that the orientation vector $\boldsymbol{\theta}_{or}$ illustrated in Fig. 2.1b points to the direction of the positive y-axis. The array orientation vectors are illustrated in Fig. 7.2a as black arrows.

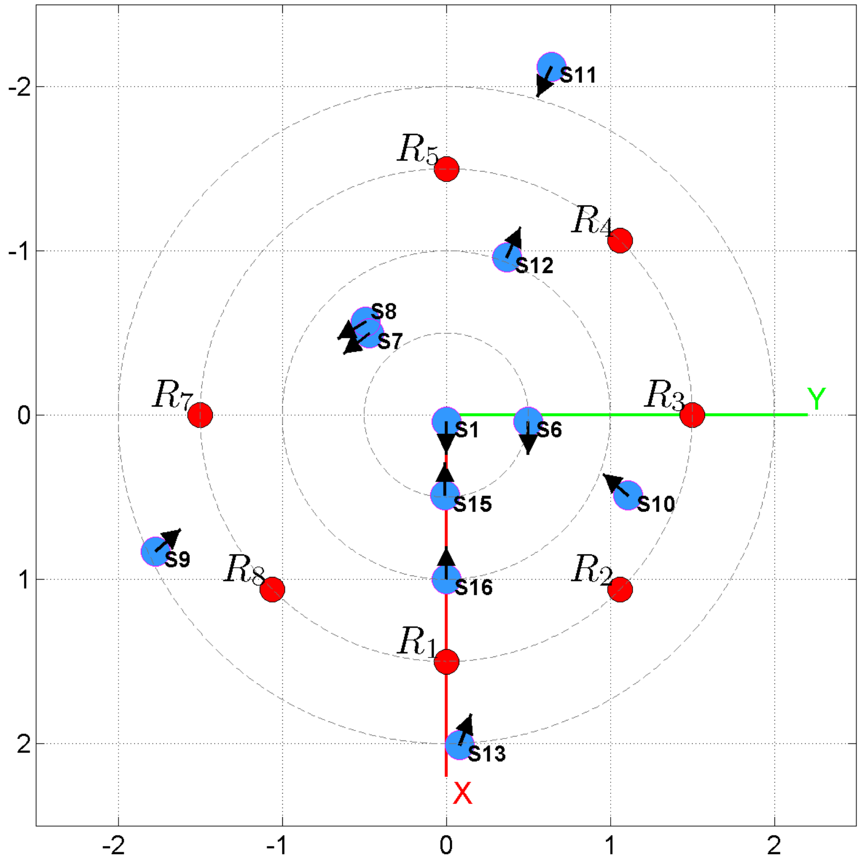



Figure 7.3: Microphone arrays R (•) and sources S (•) projected to the x-y plane.

As an acoustic source, a Genelec 8020A loudspeaker was used and placed at various source positions illustrated in Fig. 7.3. The test signals played back by the loudspeaker were exponential sweeps, which allows us to identify the channel IRs between the loudspeaker and each microphone array (see Ch. 4.3). Hence, different kinds of sources can be simulated later on by convolution of an excitation signal (e.g. female speaker) with the channel IRs.[1] Before the channel IRs are used for convolution they are windowed to a length of 2000 samples (45.4 ms), which is long enough to contain all early reflections and a fair amount of reverberation. Further the microphone and frequency-dependent gain correction factors from Ch. 4.4.4 are applied to the IRs. The sweep responses resp. IRs were measured for each source position sequentially, which allows to generate scenarios with arbitrary source combinations by superposition, assuming linearity of the propagation medium and the recording setup. The speech signals employed in the tests are illustrated in Fig. 7.4. For the single source scenarios

[1]Although the directional characteristic of the source is determined by the loudspeaker and therefore differs from the characteristic of a human speaker.

in Ch. 7.6, the signal in the first row of Fig. 7.4 is used. The multiple source scenarios described in Ch. 7.7 of 2 concurrent sources use signals in row 1,2 and scenarios of 3 concurrent sources use signals in row 1,2 and 3 of Fig. 7.4.

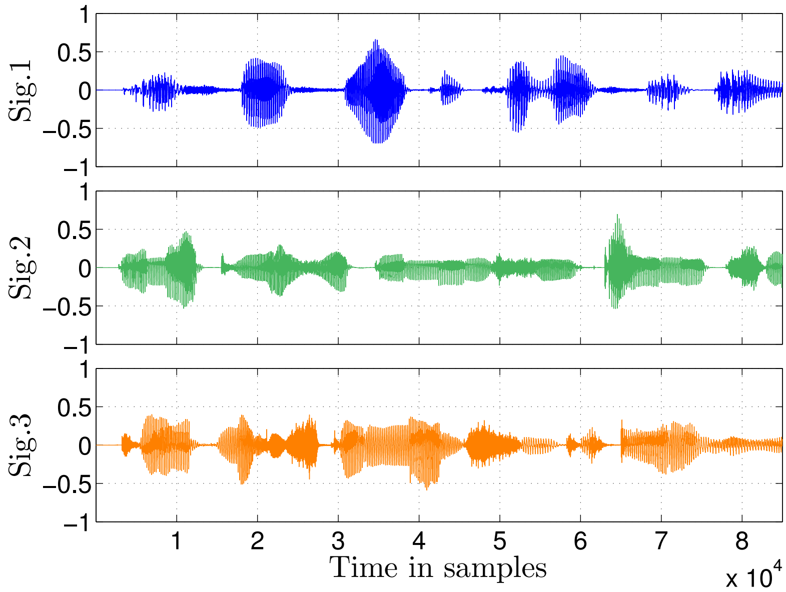



Figure 7.4: Excitation signals. Sig.1: French male speaker, Sig.2: German female speaker, Sig.3: English female speaker.

We obtain the microphone signals $\boldsymbol{s}^m[n] = (s_1^m[n], s_2^m[n], s_3^m[n], s_4^m[n])^{\mathrm{T}}$, where $m = 1, \ldots, M$ denotes the microphone array index. The following chapter describes the signal processing that is conducted on the microphone signals $\mathbf{s}_m[n]$ to evaluate the two presented localization algorithms from Ch. 3.

## 7.2 Evaluation Procedure

A flow chart describing the measurement procedure is illustrated in Fig. 7.5. All signal processing on the measured data was done offline in Mathworks Matlab. After the signals $\boldsymbol{s}^m[n]$ are obtained for each array, A to B-Format conversion (Eq. 2.6) is applied, which yields the B-Format signals $\boldsymbol{\chi}_{N_A}^m[n] = (\mathrm{W}^m[n],\ \mathrm{Y}^m[n],\ \mathrm{Z}^m[n],\ \mathrm{X}^m[n])^{\mathrm{T}}$. Despite the microphones are not strictly tetrahedral, the ideal on-axis microphone directions from Tab. 4.2 turned out to yield significantly smaller DOA estimation errors when used for B-Format encoding than the measured on-axis microphone directions given in Tab. 4.1 (Ch. 4.4.4). Therefore the ideal on-axis microphone directions are used.

The B-Format signals are then segmented into individual frames by multiplication with a Hann window of length 4096 in a way that the segments overlap by 75%. A 4096 point *Fast Fourier Transform* (FFT) is computed for each frame which yields $K = 2048$ frequency indices per frame and array. To ensure that only *valid frames*, i.e. frames that contain signal information due to an active source, are handed to the following DOA estimation procedure a signal detection algorithm is inserted. The detection algorithm discards frames, which exhibit a mean signal power below a threshold $P_{th}$. The value of $P_{th}$ is defined 15 dB above the mean signal power calculated from microphone signal segments where no source was active, i.e. only background noise was captured.

The next step is DOA estimation. Considering that DOA correction (see Ch. 6) only works up to approximately 2 kHz $\equiv$ k = 186, intensity-vector-based DOA estimation is only carried out for frequency indices k = $1, \ldots, 186$. For each frame, DOA estimates $\hat{\boldsymbol{\theta}}_k^m$ are obtained. DOA correction is then applied with the individual correction matrices from Ch. 6. The corrected DOA estimations are used to generate the 2-dimensional DOA histograms $\boldsymbol{h}_m$ described in Ch. 2.1.1. To avoid fluctuating peak positions in the DOA histograms, a third order FIR low-pass filter is implemented.

$$\tilde{\boldsymbol{h}}_{LP}^m(\eta) = \tilde{\boldsymbol{h}}^m(\eta) + \sum_{i=1}^{3} \alpha \tilde{\boldsymbol{h}}^m(\eta - i), \tag{7.1}$$

where $\eta$ denotes the frame index and $\alpha = 0.5$

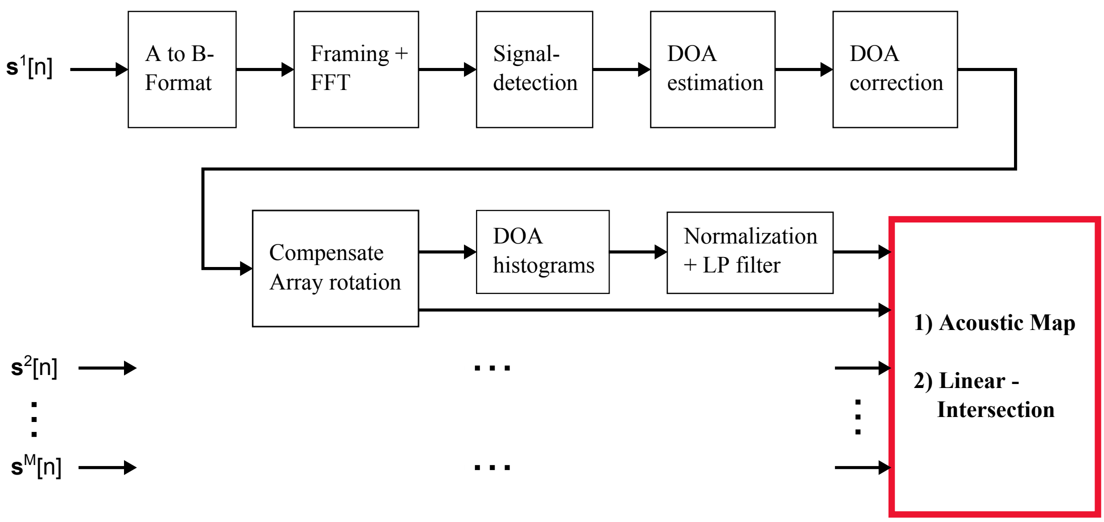



Figure 7.5: Signal flow

In a real-time wireless scenario it is favorable to compute the DOA estimations $\hat{\boldsymbol{\theta}}_k^m$ locally on each sensor node of the WiLMA system and to transmit them over the channel instead of transmitting the recorded microphone signals. By doing so, the required bandwidth and the computational load on the central unit can be reduced. The following steps; DOA correction, array rotation compensation (see Ch. 7.3), DOA histogram generation and the source localization in 3 dimensions with the acoustic map and the linear intersection algorithm can then be processed on a central unit. The acoustic map algorithm requires the DOA histograms of the $M$ arrays as input data and the linear intersection algorithm requires the corrected DOA estimations as an input and the DOA histograms for weighting purposes (see Ch. 3).

## 7.3 Correcting Microphone Array Rotation

Due to the shape of the microphone arrays it was difficult to mount optical tracking sensors to the arrays and the calibration of the array orientation vector $\boldsymbol{\theta}_{or}$ had to be done by eye. Slight errors in the alignment of the array orientation directly affect the localization performance. Therefore in this chapter we will derive a rotation axis and a rotation angle according to [Kwo98] for each microphone array, that describe potential array rotations and will be used to compensate them.

First we define the correlation matrix $\mathbf{C}_m$ of the $m^{th}$ microphone array as

$$\mathbf{C}_m = \frac{1}{N} \sum_{i=1}^{N} \hat{\boldsymbol{\theta}}_i^m \; (\boldsymbol{\theta}_i^m)^T, \tag{7.2}$$

where $\hat{\boldsymbol{\theta}}_i^m = (\hat{x}, \hat{y}, \hat{z})^T$ corresponds to the $i^{th}$ DOA estimation and $\boldsymbol{\theta}_i^m = (x, y, z)^T$ to the $i^{th}$ real source direction of microphone array $m$. N denotes the total number of considered DOA estimation/real source direction pairs. Here we use the DOA estimations of sources $S = \{1, 6, 7, 9, 10, 11, 12, 15\}$ at frequencies $f = \{320, 430, 540, 645 \text{ Hz}\}$, which yields $N = 32$ data pairs for each microphone array. DOA estimation is performed with windowed IRs of 200 samples length. The array index $m$ is omitted from now on for the sake of better readability. The singular value decomposition of matrix $\mathbf{C}$ yields

$$\mathbf{C} = \mathbf{U}\mathbf{W}\mathbf{V}^T, \tag{7.3}$$

with orthogonal matrices $\mathbf{U}$,$\mathbf{V}$ and diagonal matrix $\mathbf{W}$. A rotation matrix $\mathbf{R}$, which satisfies

$$\boldsymbol{\theta}_i = \mathbf{R}\ \hat{\boldsymbol{\theta}}_i \tag{7.4}$$

in a least-squares sense is obtained by [Kwo98]

$$\mathbf{R} = \mathbf{U} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \det(\mathbf{U}\mathbf{V}^T) \end{pmatrix} \mathbf{V}^T. \tag{7.5}$$

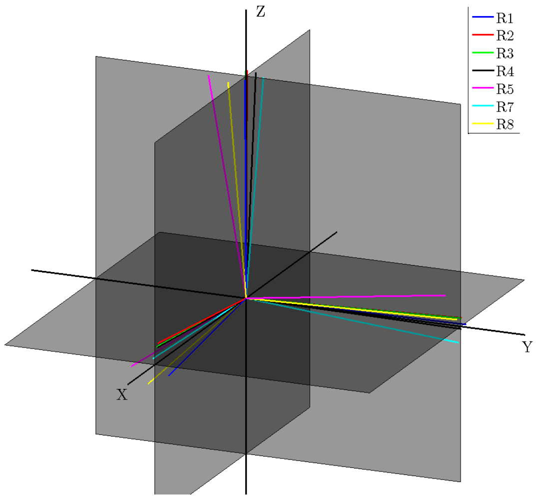
From rotation matrix $\mathbf{R}$, the rotation axis $\mathbf{a}$ (a unit vector) and the angle of rotation $\alpha_{rot}$ in a right hand sense around that axis can be derived as

$$\alpha_{rot} = \arccos\left(\frac{\mathrm{trace}(\mathbf{R}) - 1}{2}\right) \tag{7.6}$$

and

$$\mathbf{a} = \frac{1}{2\sin(\alpha_{rot})} \begin{pmatrix} \mathbf{R}(3,2) - \mathbf{R}(2,3) \\ \mathbf{R}(1,3) - \mathbf{R}(3,1) \\ \mathbf{R}(2,1) - \mathbf{R}(1,2) \end{pmatrix}. \tag{7.7}$$

$\mathbf{R}$ is a $3 \times 3$ matrix and its 3 columns indicate the basis vectors of the arrays rotated coordinate system. The base vectors are illustrated in Fig. 7.6a for all 7 microphone arrays and $\mathbf{a}$, $\alpha_{rot}$ are given in Tab. 7.6b.



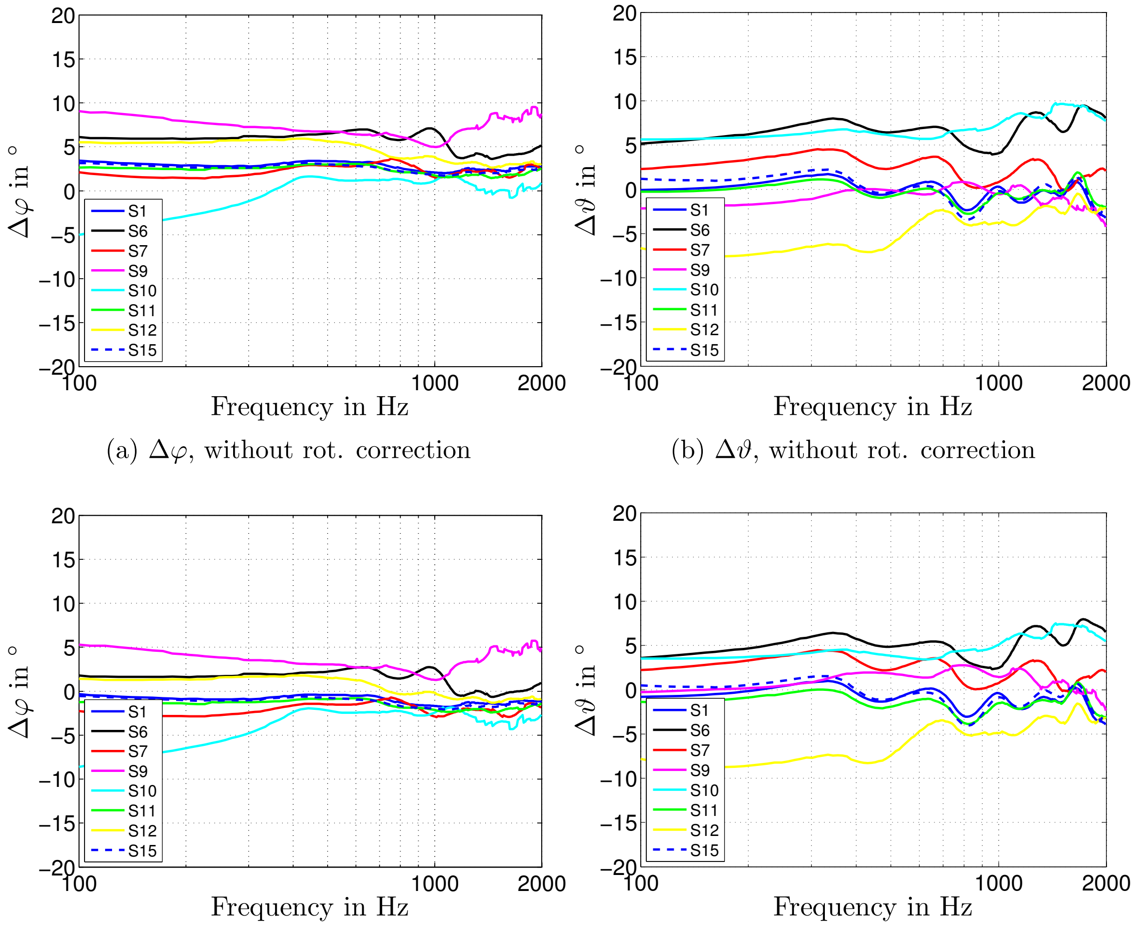

(a) Array rotations

| Array | $a_x$ | $a_y$ | $a_z$ | $\alpha_{rot}$ in ° |
|---|---|---|---|---|
| $R_1$ | -0.15 | 0.56 | 0.82 | 4.42 |
| $R_2$ | 0.34 | -0.93 | 0.12 | 5.9 |
| $R_3$ | 0.41 | -0.91 | 0.08 | 4.96 |
| $R_4$ | -0.17 | -0.98 | 0.1 | 4.48 |
| $R_5$ | 0.83 | 0.07 | -0.55 | 12 |
| $R_7$ | -0.93 | -0.34 | -0.14 | 4.32 |
| $R_8$ | 0.41 | 0.85 | -0.34 | 6.17 |

(b) Rotation axis and angle

Figure 7.6: Array rotations

Fig. 7.7 illustrates the effect of the rotation correction with the axes and angles from Tab. 7.6b on the DOA estimation errors for various source positions. Rotation correction is applied after DOA correction as depicted in the flow chart from Fig. 7.5 and manages to improve the DOA estimation performance for most source directions.




(a) $\Delta\varphi$, without rot. correction

(b) $\Delta\vartheta$, without rot. correction

(c) $\Delta\varphi$, with rot. correction

(d) $\Delta\vartheta$, with rot. correction

Figure 7.7: DOA estimation error after DOA correction (Ch. 6) in azimuth ($\Delta\varphi$) and zenith ($\Delta\vartheta$) in $^\circ$ of microphone array 1 for multiple static single sources, (a,b) before and (c,d) after rotation correction with the rot. axis and angle from Tab. 7.6b. (IR length = 200 samples)

## 7.4 Error Metrics

To evaluate the performance of the two localization algorithms, various error metrics are introduced in this chapter. In an acoustic scenario with active sources $S_i$, $q = i, \ldots, I$ the localization error of the $i^{th}$ source is defined as

$$e_i(\eta) = \| \underbrace{\mathbf{s}_i(\eta) - \hat{\mathbf{s}}_i(\eta)}_{\Delta\mathbf{s}_i(\eta)} \|, \tag{7.8}$$

where $\mathbf{s}_i(\eta)$ denotes the true and $\hat{\mathbf{s}}_i(\eta)$ the estimated source position, $\Delta\mathbf{s}_i(\eta)$ is the difference vector, $\eta$ is the frame index and $\|\cdot\|$ is the euclidean norm. An error metric independent of $\eta$ is the *mean absolute error* (MAE) given by [Fre10]

$$MAE = \underset{\eta,i}{\text{mean}}\{\|e_q(\eta)\|\} \tag{7.9}$$

where $\underset{\eta,i}{\text{mean}}\{\cdot\}$ denotes the mean over all frames $\eta$ and sources $S_i$.
Further, let us denote the number of source position estimates with $N_c$, where the localization errors $e_i$, $i = 1, \ldots, I$ of all active sources are smaller than the upper error bound $\Delta$ and the total number of estimates with $N_t$. Then the *frame accuracy* is defined as [Fre10, BOS10]

$$Acc_\Delta = \frac{N_c}{N_t} \cdot 100. \tag{7.10}$$

Note that the data considered for the error metrics are all valid frames, i.e. also frames where the number of source positions found by the algorithms $I_{est}$ is smaller than the number of real sources $I$. This can have two reasons; the algorithm's *number of source estimation* procedure only detected $I_{est} < I$ sources or source estimations were removed. A source estimation is removed if the distance to the preceding estimation for the actual frame is smaller than 1 m. This criterion is necessary since the algorithms oftentimes generate source estimations that relate to the same true source. E.g. a reason for this is that in the acoustic map's de-emphasis process the peak corresponding to the first source estimation is not removed completely from the histograms $\boldsymbol{h}^m$ and the residues cause a second source estimation. As a downside this procedure also limits the resolution of separable neighboring sources to 1 m.
Estimated source locations $\hat{\mathbf{s}}_i(n)$ are assigned to the corresponding true source positions $\mathbf{s}_i(n)$ by a minimum distance criterion, since source identification is not a part of this work [BOS10].

## 7.5 Different SNR Conditions

By adding white Gaussian noise to each microphone signal acoustic scenarios with increased surrounding diffuse noise can be simulated. Hence we can evaluate the performance of the two localization algorithms under noisy conditions. A noisy signal $\boldsymbol{s}[n]$ with a desired SNR between its clean $\boldsymbol{s}_s[n]$ and noise part $\boldsymbol{\nu}_d[n]$ can be simulated by [Fre10]

$$\boldsymbol{s}[n] = \boldsymbol{s}_s[n] + \underbrace{G_{SNR} \cdot \boldsymbol{\nu}[n]}_{\boldsymbol{\nu}_d[n]}, \tag{7.11}$$

where

$$G_{SNR} = \sqrt{\frac{P_d}{P_\nu}}, \tag{7.12}$$

$$P_d = P_s \; 10^{-SNR/10}. \tag{7.13}$$

$\boldsymbol{\nu}[n]$ is white Gaussian noise with signal power $P_\nu$ and $P_s$, $P_d$ denote the signal power of $\boldsymbol{s}_s[n]$ and the desired noise component $\boldsymbol{\nu}_d[n]$. Due to background noise that is captured by the microphones, we will never obtain a completely clean signal $\boldsymbol{s}_s[n]$, which is why the upper limit for the desired SNR is given by the SNR of $\boldsymbol{s}_s[n]$.

For the acoustic scenario of single active source $S_1$ the clean signals recorded by microphone array $R_1$, which is positioned in the radiation direction of $S_1$, exhibit the following SNR values:

- Mic.2 (facing $S_1$): 36 dB
- Mic.3 (opposed to $S_1$): 10 dB

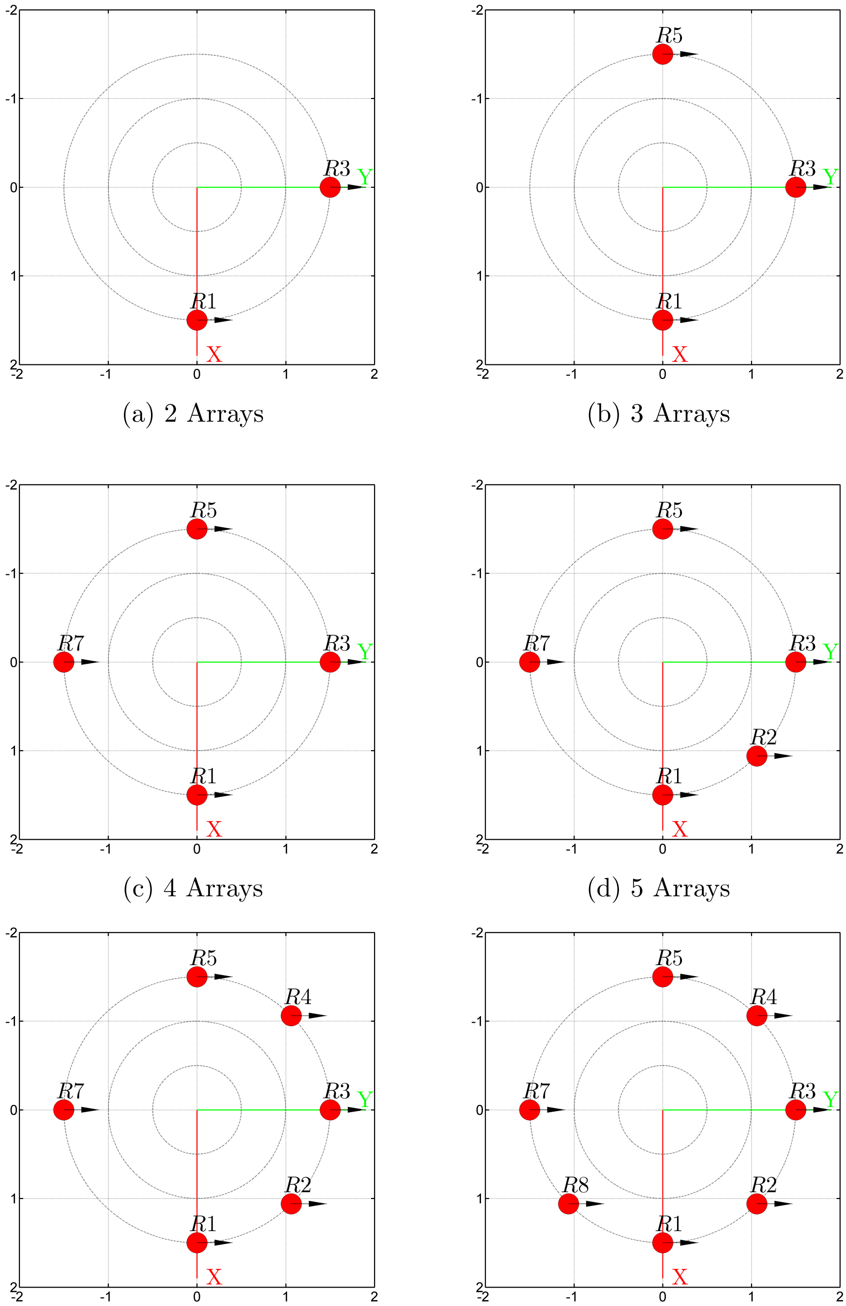
Microphone array $R_5$ positioned in the rear of $S_1$ yields:

- Mic.3 (facing $S_1$): 31 dB
- Mic.2 (opposed to $S_1$): 5 dB

To evaluate the two localization algorithms 2 different SNR scenarios will be used. In the first scenario the clean microphone signals are used (no white noise added). From now we will use the mean value SNR = 24 dB of all 7 Array's (source $S_1$) to refer to the first scenario. By adding white noise (Eq. 7.11) a second scenario is simulated with SNR = 10 dB.

## 7.6 Static Single Source Experiments

In this chapter we will evaluate the ability of the acoustic map and linear intersection algorithm to localize the position of an active single source in 3 dimensions. For this purpose a loudspeaker is placed at the positions $S = \{1, 6, 7, 9, 10, 11, 12, 15\}$ shown in Fig. 7.3. The windowed channel IRs are then convolved with the excitation signal of a French male speaker ($1^{st}$ row Fig. 7.4). The resulting 1.93 $s$ long microphone array signals are segmented into 80 frames, i.e. a new frame is generated every 23.2 ms



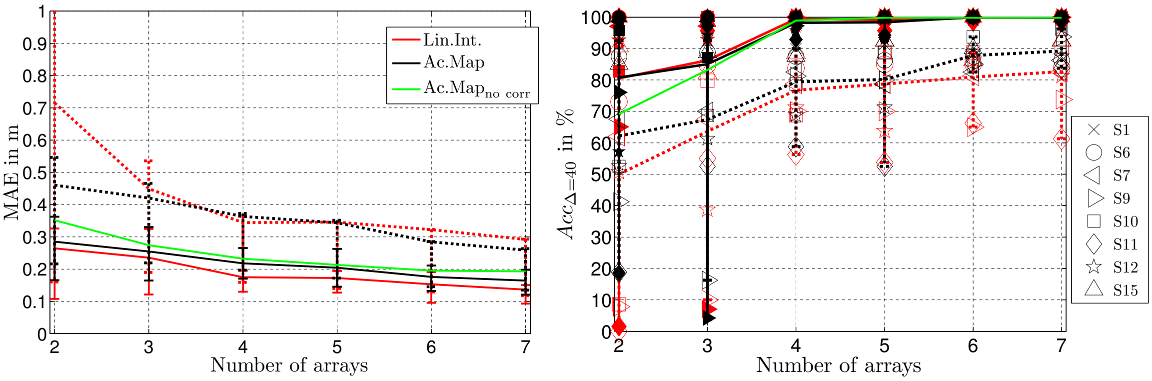

(a) 2 Arrays (b) 3 Arrays (c) 4 Arrays (d) 5 Arrays (e) 6 Arrays (f) 7 Arrays

Figure 7.8: Constellations using different numbers of microphone arrays.

$\equiv$ 1024 samples.

To investigate the influence of the number of microphone arrays on the localization performance, different array combinations were used by successively increasing the number of used arrays as shown in Fig. 7.8, starting from 2 to a maximum number of 7 arrays.

Fig. 7.9a shows the MAE from Eq. 7.9 for both acoustic map and linear intersection algorithm and two different SNR conditions with 24 dB and 10 dB SNR. Observing the 24 dB SNR scenario the MAE curves of the 2 algorithms stay below 0.3 m for all numbers of active arrays. Increasing the number of active arrays, the localization performance of both algorithms improves, as the MAE steadily decreases to a value of approximately 0.15 m for 7 active arrays. It is noticeable that the linear intersection algorithm performs slightly better than the acoustic map algorithm. Compared to the 24 dB scenario the 10 dB SNR scenario shows an MAE increase of 0.15 m. Here the acoustic map algorithm performs better.



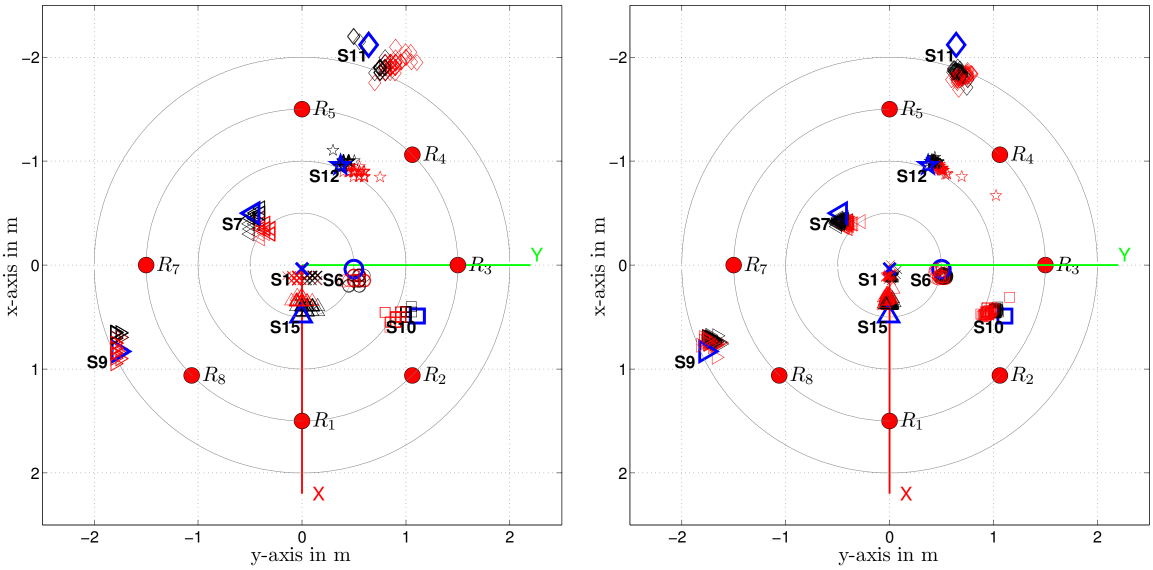

(a) MAE

(b) $Acc_{\Delta=40}$

Figure 7.9: Comparison between acoustic map and linear intersection algorithm in terms of error metrics MAE and $Acc_{\Delta=40}$ for single static source positions $S = \{1, 6, 7, 9, 10, 11, 12, 15\}$ and a SNR of 24 dB (solid lines) and 10 dB (dotted lines). The error bars in (a) mark the first and third quantiles. Markers in (b) show the $Acc_{\Delta=40}$ values computed individually for each source position and the error bars show the range of these values. Valid frames/total frames, 24 dB SNR: 70/80, 10 dB SNR: 80/80.

The accuracy rate $Acc_{\Delta=40}$ (Eq. 7.10) values are illustrated in Fig. 7.9b. Here both algorithms perform almost equally for the 24 dB scenario. In analogy to Fig. 7.9a the acoustic map algorithm has the advantage in the noisy 10 dB SNR case. The highest increase in performance can be observed going from 2 to 3 and from 3 to 4 active microphone arrays. This exemplifies the fact, that localization performance highly depends on the array positioning and array coverage in the volume of possible source locations. For 3 active arrays the markers in Fig. 7.9b indicate that the arrays $R_1$,$R_3$,$R_5$ have problems with the localization of source position $S_9$. Adding array

$R_7$, which is positioned in close vicinity to $S_9$, leads to a drastic improvement in localization accuracy.
To evaluate the influence of the DOA correction procedure from Ch. 6, Fig. 7.9a and Fig. 7.9b also shows the MAE and $Acc_{\Delta=40}$ curve of the acoustic map algorithm, where no DOA correction was applied to the DOA estimation data before handing it to the algorithm. The curves show that the localization performance without DOA correction is already quite accurate. The DOA correction manages to slightly decrease the MAE values by approximately $1 \ldots 7$ cm depending on the number of arrays and might be interesting for applications, where accuracy is critical.

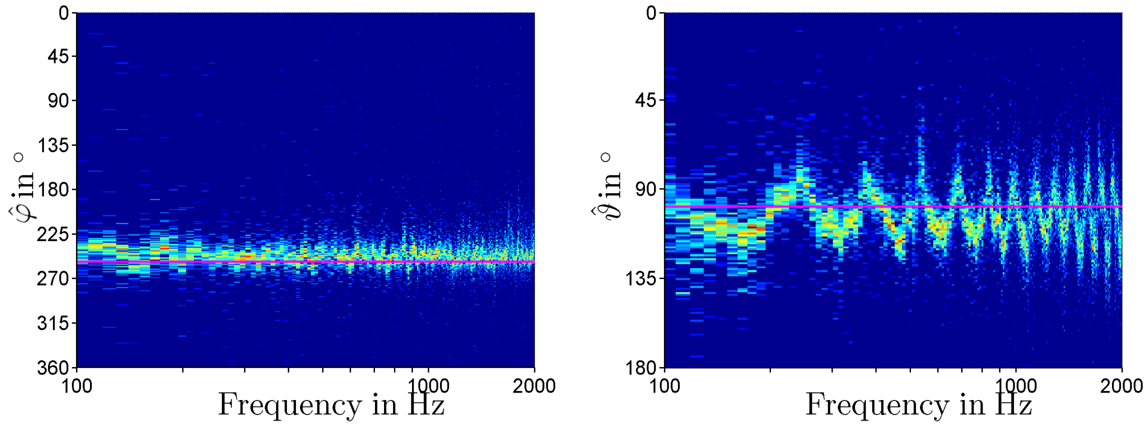



(a) Acoustic map (b) Linear intersection

Figure 7.10: Localization estimates projected onto the x-y plane, of 50 frames for each single static source positions $S = \{1, 6, 7, 9, 10, 11, 12, 15\}$ for (a) acoustic map and (b) linear intersection algorithm. Red markers denote localization estimates generated with microphone arrays $R = \{1, 3, 5, 7\}$, black markers with arrays $R = \{1, 2, 3, 4, 5, 7, 8\}$ and blue markers correspond to the true source locations.

An interesting observation can be made by looking at the behavior of the DOA estimations in zenith $\hat{\vartheta}$ over frequency, exemplarily illustrated in Fig. 7.11b for microphone array $R_7$ and single source $S_9$. $\hat{\vartheta}$ periodically fluctuates around the true zenith angle in a zig-zag pattern with a period of approximately 150 Hz. This effect is most likely caused by strong floor reflections. If we only consider single-source scenarios, a simple smoothing over frequency might improve the quality of the DOA estimations significantly.[2]

[2]For multiple source scenarios smoothing is not feasible, since DOA estimation angles jump over




(a) $\hat{\varphi}$ histogram

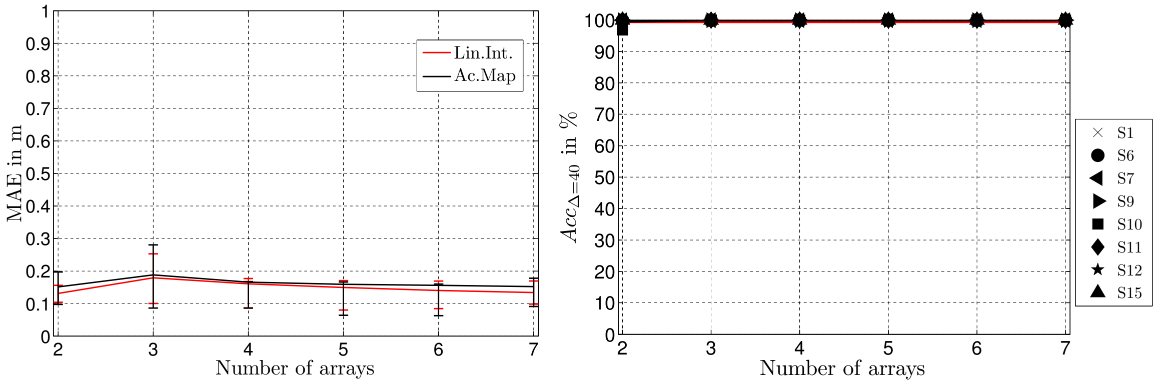
(b) $\hat{\vartheta}$ histogram

Figure 7.11: Distribution of DOA estimation data $\hat{\varphi}_{\mathrm{k}}$, $\hat{\vartheta}_{\mathrm{k}}$ of all 80 frames illustrated for each frequency index k from 100 . . . 2000 Hz for microphone array $R_7$ and single source $S_9$. The magenta colored line indicates the true DOA angle.

**Source Localization without Room Reflections:** To gain insight in how the two algorithms behave in a completely anechoic room, where no signal distortion due to room reflections is present, the IRs are windowed to a length of 150 Samples. The corresponding MAE and $Acc_{\Delta=40}$ error metrics are shown in Fig. 7.12 and the localization estimates are illustrated in Fig. 7.13. Fig. 7.12a indicates, that even without any room reflections the minimum achievable MAE adds up to 0.15 m, whereas the accuracy rate $Acc_{\Delta=40}$ shows a clear improvement compared to Fig. 7.9b with a value of approximately 100% for all array numbers.

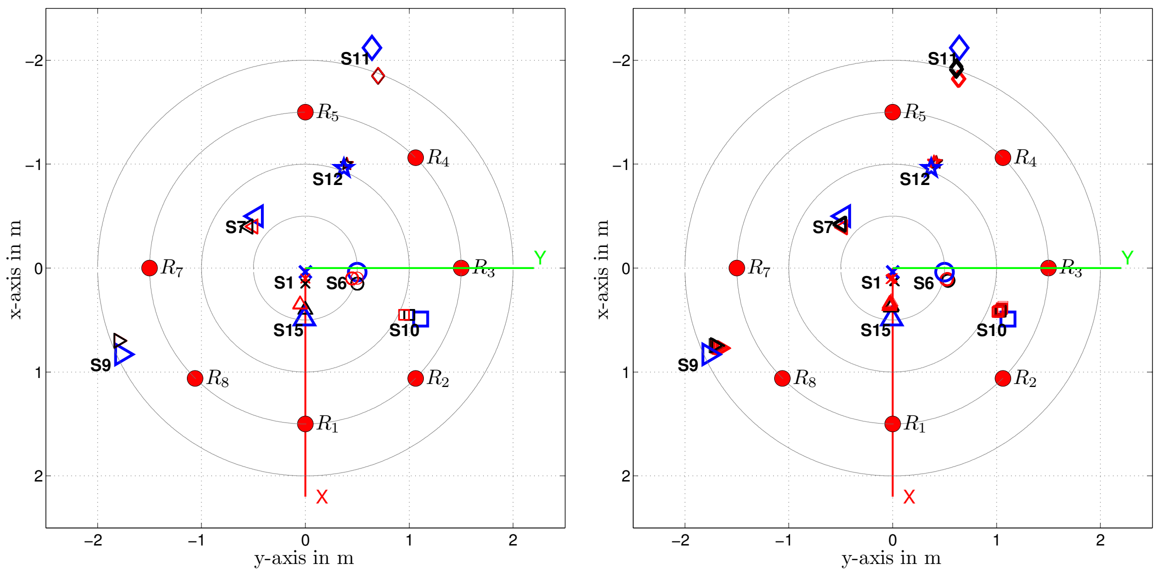
Under the assumption of a fully functional DOA correction, the MAE should ideally exhibit values close to zero in an anechoic scenario. A possible reason that the MAE values in Fig. 7.12a do not fall below 0.15 m might be measurement errors e.g. rotated array positioning that can not be corrected or derivations of the loudspeaker semicircle used in measurement 1 (Ch. 4) from an ideal semicircle. These errors prevent the DOA correction derived from the first measurement to be fully applicable to the DOA estimates of measurement 2 (Ch. 7). Further Fig. 7.12a indicates, that microphone array 5 generates imprecise DOA estimates since the MAE value increases from 2 to 3 active microphone arrays.

frequency from 1 source to another and smoothing would wrongly smooth these important jumps.




(a) MAE (b) $Acc_{\Delta=40}$

Figure 7.12: Comparison between acoustic map and linear intersection algorithm in terms of error metrics MAE and $Acc_{\Delta=40}$ for single static source positions $S = \{1, 6, 7, 9, 10, 11, 12, 15\}$. The error bars in (a) mark the first and third quantiles. Markers in (b) show the $Acc_{\Delta=40}$ values computed individually for each source position and the error bars show the range of these values. Valid frames/total frames, 24 dB SNR: 61/80.

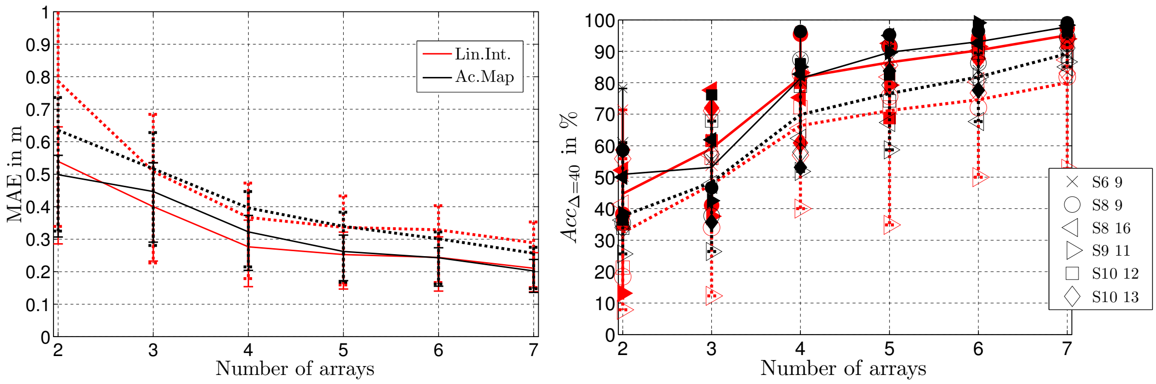



(a) acoustic map (b) Linear intersection

Figure 7.13: Localization estimates projected onto the x-y plane, of 50 frames for each single static source positions $S = \{1, 6, 7, 9, 10, 11, 12, 15\}$ for (a) acoustic map and (b) linear intersection algorithm. Red markers denote localization estimates generated with microphone arrays $R = \{1, 3, 5, 7\}$, black markers with arrays $R = \{1, 2, 3, 4, 5, 7, 8\}$ and blue markers correspond to the true source locations.

## 7.7 Static Multiple Source Experiments

In the first part of this chapter, scenarios with 2 active sources and in the second part scenarios with 3 active sources are investigated. For the scenario with 2 active sources the source pairs $S = \{[6\ 9], [8\ 9], [8\ 16], [9\ 11], [10\ 12], [10\ 13]\}$ (Fig. 7.3) are used. The IRs of the first source of each pair are convolved with the recorded sample of a French male speaker ($1^{st}$ row Fig. 7.4) and the IRs of the second source with the recorded sample of a German female speaker ($2^{nd}$ row Fig. 7.4). In a multiple source scenario with $I$ active sources, the acoustic map algorithm successively estimates the source positions until the amplitude in the acoustic map corresponding to the found source position lies below a threshold $\mathcal{L}_{Th}$. Estimating the number of active sources based on $\mathcal{L}_{Th}$ turns out to be error-prone. To ensure a fair comparison between the two localization algorithms, the information on the number of active sources obtained via the gap-statistic in the linear intersection algorithm is also provided to the acoustic map algorithm.




(a) MAE (b) $Acc_{\Delta=40}$

Figure 7.14: Comparison between acoustic map and linear intersection algorithm in terms of error metrics MAE and $Acc_{\Delta=40}$ for 2 simultaneously active static sources $S = \{[6\ 9], [8\ 9], [8\ 16], [9\ 11], [10\ 12], [10\ 13]\}$ and a SNR of 24 dB (solid lines) and 10 dB (dotted lines). The error bars in (a) mark the first and third quantiles. Markers in (b) show the $Acc_{\Delta=40}$ values computed individually for each source position and the error bars show the range of these values. Valid frames/Total frames: 80/80 (24 dB SNR), 80/80 (10 dB SNR). Frames with 2 est./Frames with 1 est.: 1915/920 (24 dB SNR), 1933/913 (10 dB SNR).

The MAE and $Acc_{\Delta=40}$ illustrated in Fig. 7.14 significantly improve with an increasing number of used microphone arrays. In the 24 dB SNR case, from around 50 . . . 60 cm MAE and 45 . . . 55% $Acc_{\Delta=40}$ with 2 used arrays to 20 cm MAE and 90 . . . 100% $Acc_{\Delta=40}$ with 7 used arrays. Decreasing the SNR to 10 dB yields an increase in MAE of approx. 10 cm for both algorithms. The $Acc_{\Delta=40}$ decreases by approx. 15% for the linear intersection and 10% for the acoustic map algorithm, i.e. the acoustic map algorithm performs slightly better in a noisy scenario.

In Fig. 7.15 the localization error of the source position estimates corresponding to the first and the second peak in the acoustic map is illustrated over time frames $\eta$ exemplarily for source positions $S = \{6\ 9\}$.

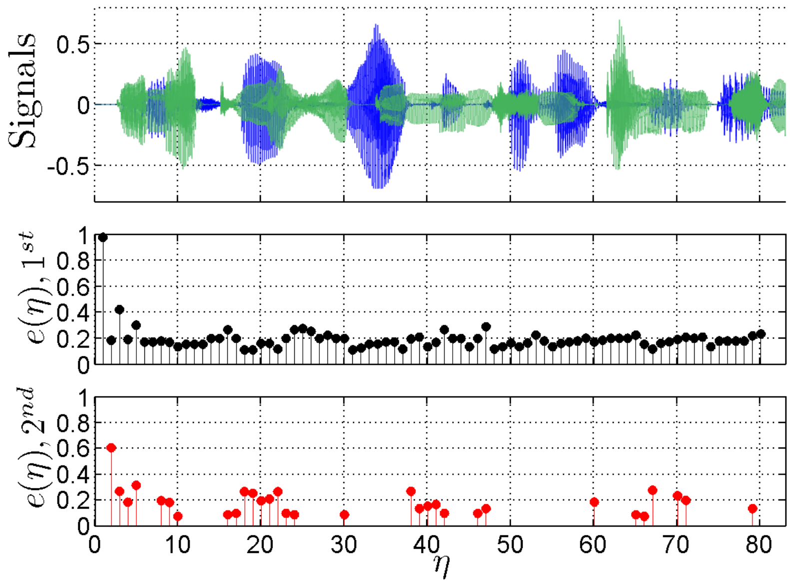
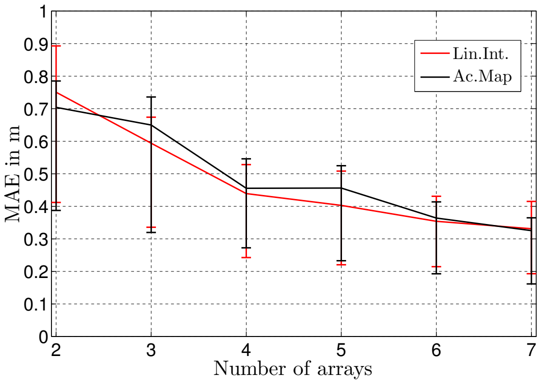
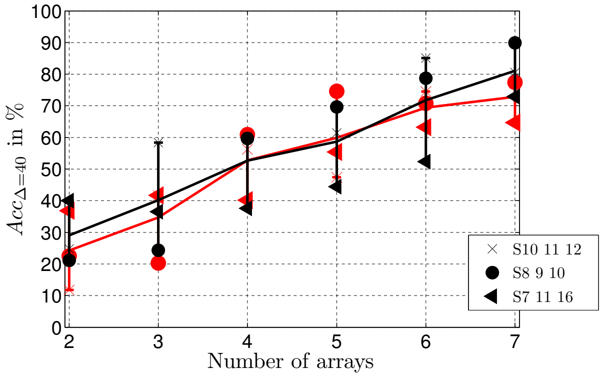



Figure 7.15: Error $e(\eta)$ of the $1^{st}$ (highest peak in ac.map) and the $2^{nd}$ source position estimate (highest peak in ac.map after de-emphasis) found by the acoustic map algorithm for 2 active speakers at positions $S = \{6\ 9\}$ over time frames $\eta$. Excitation signals Sig.1 (French male speaker), Sig.2 (German female speaker) from Fig. 7.4 are illustrated overlaid in the first row.

For the three source scenario the source pairs $S = \{[10\ 11\ 12], [8\ 9\ 10], [7\ 11\ 16]\}$ (Fig. 7.3) are used. The IRs of the first source of each triplet are convolved with the recorded sample of a French male speaker ($1^{st}$ row Fig. 7.4), the IRs of the second source with the recorded sample of a German female speaker ($2^{nd}$ row Fig. 7.4) and the IRs of the third source with the recorded sample of an English female speaker ($3^{rd}$ row Fig. 7.4). MAE and $Acc_{\Delta=40}$ values are shown in Fig. 7.16.




(a) MAE




(b) $Acc_{\Delta=40}$

Figure 7.16: Comparison between acoustic map and linear intersection algorithm in terms of error metrics MAE and $Acc_{\Delta=40}$ for 3 simultaneously active static sources $S = \{[10\ 11\ 12], [8\ 9\ 10], [7\ 11\ 16]\}$ and a SNR of 24 dB (solid lines) and 10 dB (dotted lines). The error bars in (a) mark the first and third quantiles. Markers in (b) show the $Acc_{\Delta=40}$ values computed individually for each source position and the error bars show the range of these values. Valid frames/Total frames: 80/80 (24 dB SNR), 80/80 (10 dB SNR). Frames with 3 est./2 est./1 est.: 768/474/178 (24 dB SNR).

# Chapter 8

# Conclusion and Outlook

In this work, two different ASL algorithms were developed, namely the acoustic map and linear intersection algorithm, that allow to localize multiple simultaneously active sound sources in 3D space by processing the DOA estimation data from multiple distributed tetrahedral microphone arrays.

To evaluate the DOA estimation performance of the individual Oktava 4D-Ambient tetrahedral arrays, a first measurement was conducted. It was shown, that the DOA error in azimuth and zenith averaged over all measured source locations stays below approximately 5° in the frequency range up to 2 kHz and the directivity patterns of all 4 cardioid microphones constituting a tetrahedral array show a high degree of symmetry, which is desirable for acoustic scene analysis applications like ASL.[1]

In a second measurement, the two ASL algorithms are evaluated. The localization performance of both algorithms increases with the number of used tetrahedral microphone arrays in single and multiple source scenarios. In a scenario where the maximum number of 7 available microphone arrays are used with an SNR of 24 dB, a mean absolute error of 13 . . . 18 cm could be reached for 1 active source and 20 cm for 2 active sources. Comparing the two algorithms, it could be observed that the linear intersection algorithm performs slightly better in scenarios with 24 dB SNR and the acoustic map had the advantage in noisy scenarios with 10 dB SNR.

As the accuracy of the estimated source positions in 3D space depends on the DOA estimation performance of the individual tetrahedral arrays, frequency dependent DOA correction matrices were derived and applied to the individual arrays, which

[1] This is true for a completely dry room, i.e. the measured IRs that were used to evaluate symmetry of the directivity patterns and the DOA performance were cut to a length of 150 samples.

leads to a reduction of the mean absolute error of approximately $1 \ldots 7$ cm depending on the number of tetrahedral arrays.

**Outlook:** A real-time implementation of the offline algorithm is a natural next step. It would be an exciting future accomplishment to fix one problem that occured during the practical experiments. It was a big challenge in this work to identify potential rotation errors in the positioning of the individual tetrahedral arrays. Hence, for a hands-on ASL system, an algorithm that accurately localizes the position and rotation of each tetrahedral array could offer an invaluable assistance.

An estimate on the number of active sources is needed by both developed ASL algorithms. The implemented procedures in this work that generate this estimate still require optimization, especially in the case of the acoustic map algorithm.

# Appendix A

# Theoretical Background

## A.1 Spherical Coordinates

Since the microphone arrays used in this work are spherical arrays, it is convenient to use a spherical coordinate system. Fig. A.1 shows the used conventions.

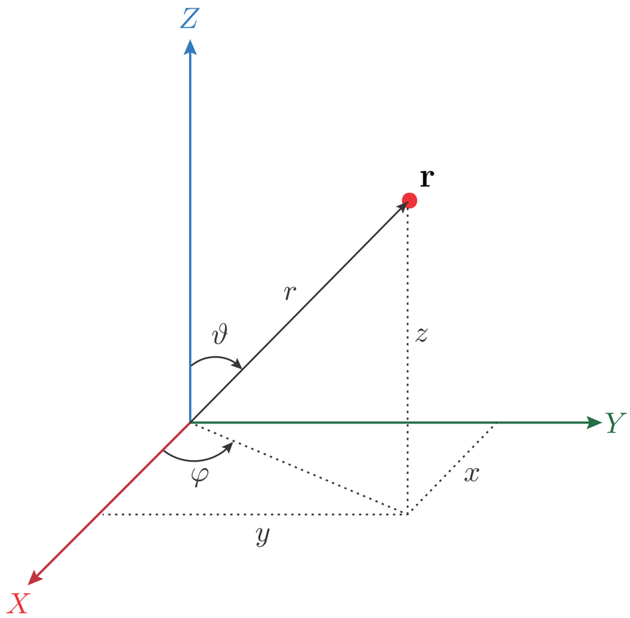



Figure A.1: Spherical coordinate system [Pau13]

A vector $\mathbf{r} \in \mathbb{R}^3$ is described by

$$\mathbf{r} = \begin{pmatrix} x \\ y \\ z \end{pmatrix} = r \begin{pmatrix} \cos(\varphi)\sin(\vartheta) \\ \sin(\varphi)\sin(\vartheta) \\ \cos(\vartheta) \end{pmatrix} = r\boldsymbol{\theta}, \tag{A.1}$$

with

$$
\begin{aligned}
r &= \sqrt{x^2 + y^2 + z^2} \\
\varphi &= arc\tan\left(\frac{x}{y}\right) \\
\vartheta &= arc\cos\left(\frac{z}{r}\right)
\end{aligned}
\tag{A.2}
$$

where $r \in [0, \infty]$ is the radius, $\varphi \in [0, 2\pi]$ the azimuth angle and $\vartheta \in [0, \pi]$ the zenith angle. $\boldsymbol{\theta}$ denotes the position $(\varphi, \vartheta)$ on the unit sphere.

## A.2 Solving the Wave Equation

If we assume linearity and zero viscosity of the propagation medium, the pressure in a source free sound field can be described by the homogeneous linear and lossless wave equation, which is defined as

$$
\Delta p(\mathbf{r}, t) - \frac{1}{c^2}\frac{\partial^2 p(\mathbf{r}, t)}{\partial t^2} = 0. \tag{A.3}
$$

Transforming Eq. A.3 into the frequency domain leads to the so called Helmholtz equation

$$
(\Delta + k^2)p(\mathbf{r}, \omega) = 0. \tag{A.4}
$$

Since a solution in spherical coordinates is desired, the Laplacian from Eq.A.3 is expressed in spherical coordinates and Eq.A.3 is solved through a product-ansatz and a separation of variables

$$
p(\mathbf{r}, \omega) = R(kr)\Phi(\varphi)\Theta(\vartheta), \tag{A.5}
$$

Inserting Eq. A.5 into Eq. A.4 separates Eq. A.4 into 3 ordinary known differential equations in $r$, $\varphi$ and $\vartheta$. The 3 equations are solved independently and a radial solution $R(kr)$ and 2 angular solutions $\Phi(\varphi)$, $\Theta(\vartheta)$ are obtained. Eq. A.6, Eq. A.7 and Eq. A.9 show the solutions for the 3 ordinary differential equations, where $n, m$ are separation constants which result from the separation of variables.

**Solution in $\varphi$:** The ordinary differential equation in $\varphi$ can be solved with

$$
\phi(\varphi) = \left\{\begin{matrix} \sin(m\varphi), \cos(m\varphi) & \text{real valued} \\ e^{\pm jm\varphi} & \text{complex.} \end{matrix}\right\} \tag{A.6}
$$

Both the real valued or the complex solution is possible, depending on which of them is used in the complete solution shown later in Eq. A.11, we get the so called real valued spherical harmonics or the complex spherical harmonics. For periodicity and continuity of $\phi(\varphi)$, $m$ has to be an integer number, $m \in \mathbb{Z}$ [Wil99].

**Physical Solution in $\vartheta$:** The solutions to the ordinary differential equation in $\vartheta$ are Legendre functions of the first and second kinds

$$\Theta(\vartheta) = \Theta_1 P_n^m(\cos\ \vartheta) + \Theta_2 Q_n^m(\cos\ \vartheta). \tag{A.7}$$

Due to the fact, that the Legendre functions of the second kind $Q_n^m$ possess singularities at positions $\cos(\vartheta) = \pm 1$ they are not physically reasonable and therefor discarded ($\Theta_2 = 0$). Singularities also occur for $P_n^m$ with $\cos(\vartheta) = 1$ unless we define $n$ to have only integer values [Wil99]. We get

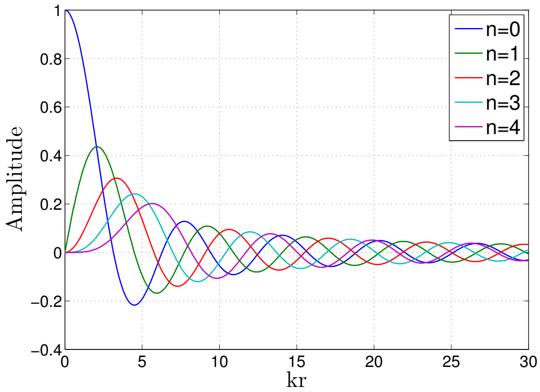
$$\Theta(\vartheta) = \Theta_1 P_n^m(\cos\ \vartheta), \qquad for n \in \mathbb{N}. \tag{A.8}$$

**Physical Solution in $r$:** The ordinary radial differential equation is solved by [Wil99]

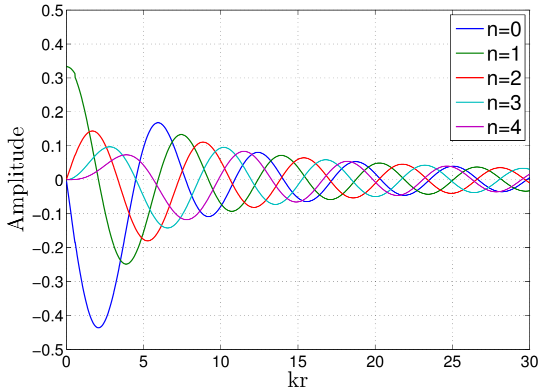
$$R_n(r) = R_1 j_n(kr) + R_2 h_n^{(2)}(kr), \tag{A.9}$$

where $j_n(kr)$ are the spherical bessel functions of the first, $h_n^{(2)}(kr)$ are the complex spherical hankel functions of second kind and $R_1, R_2$ are complex constants. Fig. A.2 shows $j_n(kr)$ and its derivative $j_n'(kr)$ for orders $n = 0, \ldots, 4$ and fig. A.3 shows the magnitude of $h_n^{(2)}(kr)$ and its derivative $h_n^{(2)\prime}$ for orders $n = 0, \ldots, 6$.

Depending on the physical problem, different parts of the solutions listed in Eq. A.9 are physically meaningful. To describe an irradiating sound field inside a source free volume with radius $r_a$, where all sources lie outside $r > r_a$ the volume (so called *Inner problem*) the radial solution has to be regular for $r \leq r_a$, i.e. the spherical bessel functions $j_n(kr)$ are used since the spherical hankel functions possess a singularity at $r = 0$ ($R_2 = 0$ in Eq. A.9). If sources are confined inside a volume with radius $r_a$ the radiating source free soundfield outside the volume is described by the spherical hankel functions $h_n^{(2)}(kr)$ (*Outer problem*), since they fulfill the *Sommerfeld radiation condition* ($R_1 = 0$ in Eq. A.9). If radiating and irradiating sources are present the source free sound-field in between is described by both spherical bessel and hankel functions (*Mixed problem*).

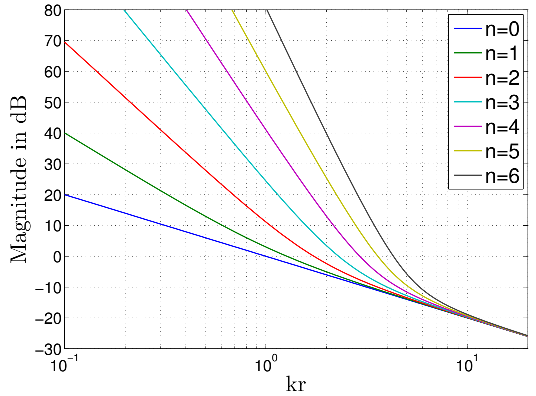



(a) $j_n(kr)$

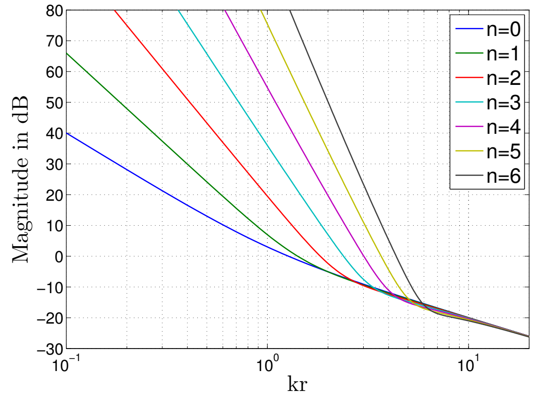



(b) $j_n'(kr)$

Figure A.2: Spherical bessel functions $j_n(kr)$ and its derivative $j_n'(kr)$ for orders $n = 0, \ldots, 4$.




(a) $h_n^{(2)}$




(b) $h_n^{(2)\prime}$

Figure A.3: Magnitude of spherical hankel functions $h_n^{(2)}(kr)$ and its derivative $h_n^{(2)\prime}$ for orders $n = 0, \ldots, 6$.

**Total solution:** By inserting the angular and radial solutions into Eq. A.5 and then summing over all modes $(n, m)$ a total solution is obtained [Wil99, Zot09, Zot10]

$$p(kr, \varphi, \vartheta) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \underbrace{\left(b_n^m j_n(kr) + c_n^m h_n^{(2)}(kr)\right)}_{\psi_n^m} Y_n^m(\varphi, \vartheta), \tag{A.10}$$

where $Y_n^m$ are the spherical harmonics, which contain the angular solutions $\Phi(\varphi)$ ans $\Theta(\vartheta)$. The coefficients $b_n^m$ describe the irradiating and $c_n^m$ the radiating part of the soundfield and $\psi_n^m$ denotes the spherical wave spectrum. The spherical harmonics are defined as

$$Y_n^m(\varphi,\vartheta) = N_n^{|m|} P_n^{|m|}(\cos\ \vartheta) \begin{cases} \cos(m\varphi) & \text{for } m \geq 0 \\ \sin(m\varphi) & \text{for } m < 0, \\ e^{\pm im\varphi} & \end{cases} \tag{A.11}$$

where $n$ stands for the order and $m$ for the degree of the particular spherical harmonic. The terms on the right side of the curly bracket corresponds to the solution in $\varphi$ of Eq. A.6. Using $\cos(m\varphi)$ and $\sin(m\varphi)$ for the solution in $\varphi$ leads to the real valued spherical harmonics. Alternatively $e^{\pm im\varphi}$ can be used to obtain the complex spherical harmonics. In the course of this work the real valued spherical harmonics will be used. $N_n^{|m|}$ is a normalization factor which arranges for the spherical harmonics to be orthonormal:

$$\int_0^{2\pi} \int_0^{\pi} Y_n^m(\varphi,\vartheta) Y_{n'}^{m'}(\varphi,\vartheta) d\varphi\ \sin(\vartheta) d\vartheta = \delta_{nn'} \delta_{mm'} \tag{A.12}$$

The real valued spherical harmonics up to order $n = 3$ are illustrated in Fig. A.4.

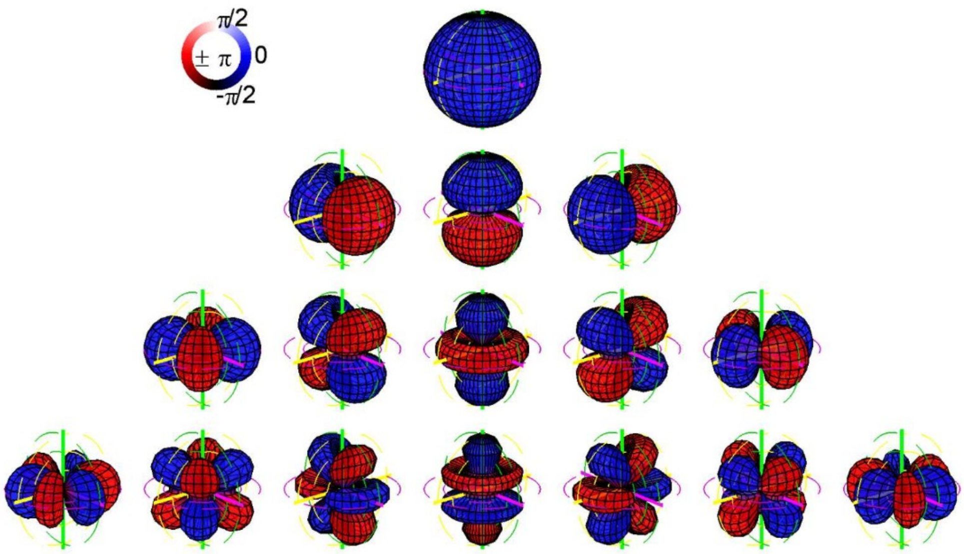



Figure A.4: Spherical harmonics $Y_n^m(\varphi,\vartheta)$ up to order $n = 3$ [Pom08].

In Eq. A.10 the coefficients $b_n^m$ and $c_n^m$ are also called wave spectrum. They contain a coefficient for each spherical harmonic independent of frequency or radius. By knowledge of $b_n^m$ and $c_n^m$ the sound-field is completely described. Multiplication with the radial functions $j_n(kr)$, $h_n^{(2)}(kr)$ yield the so called spherical wave spectrum $\psi_n^m$ of the pressure distribution $p(kR,\varphi,\vartheta)$ on the sphere of radius $R$. In contrast to the wave spectrum, the spherical wave spectrum depends on frequency and radius because it contains the radial functions. When comparing Eq. A.10 with Eq. A.14 it can be seen that $\psi_n^m$ is the Spherical harmonic transform of the pressure distribution $p(kR, \varphi, \vartheta)$ on the sphere.

From Eq. A.10 we know that the spherical wave spectrum of a pressure distribution $p(kR, \varphi, \vartheta)$ on a sphere with radius $R$ can be described by $\psi_n^m = b_n^m j_n(kR) + c_n^m h_n^{(2)}(kR)$. Since not only pressure $p(kR, \varphi, \vartheta)$ but also signals on the sphere involving pressure gradient are considered in this work, we use the notation $s(kR, \varphi, \vartheta)$ and denote the spherical wave spectrum in a more general fashion as

$$\chi_n^m = b_n^m \rho_n(kR), \tag{A.13}$$

where $b_n^m$ describes the wave spectrum and $\rho_n(kR)$ a radial propagation term. The following table lists the radial propagation terms for 3 commonly used array configurations.

| Array configuration | $\rho_n(kR)$ |
|---|---|
| Open sphere, Omni | $j_n(kR)$ |
| Rigid sphere, Omni | $j_n(kR) - \frac{j_n'(kr_0)}{h_n'^{(2)}(kr_0)} h_n^{(2)}(kR)$ |
| Open sphere, Cardioid | $\beta j_n(kR) - i(1-\beta) j_n'(kR)$ |

Table A.1: Radial propagation terms $\rho_n(kR)$ for different microphone array configurations with array radius $R$ and rigid sphere radius $r_0$ ($r_0 \leq R$).

## A.3 Spherical Harmonics Transform

Any square integrable function $f(\boldsymbol{\theta})$ on the sphere can be transformed into the spherical harmonics domain by the *Spherical harmonics transform* (SHT)

$$\gamma_n^m = \mathcal{SHT}(f(\boldsymbol{\theta})) = \int_{\boldsymbol{\theta} \in S^2} f(\boldsymbol{\theta}) Y_n^m(\boldsymbol{\theta}) d\boldsymbol{\theta}, \tag{A.14}$$

where $\gamma_n^m$ are the spherical harmonic coefficients (spherical spectrum), $Y_n^m$ are the real valued spherical harmonics described in Eq. A.11 and $\int_{\boldsymbol{\theta} \in S^2} = \int_0^{2\pi} \int_0^{\pi} \sin(\vartheta) d\vartheta \; d\varphi$. Weighting the spherical harmonics $Y_n^m$ with the corresponding coefficients $\gamma_n^m$ and then summing over all $n, m$ yields the *Inverse spherical harmonic transform* (ISHT)

$$f(\boldsymbol{\theta}) = \mathcal{ISHT}(\gamma_n^m) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \gamma_n^m Y_n^m(\boldsymbol{\theta}), \tag{A.15}$$

## A.4 Discrete Spherical Harmonics Transform

In practice a function $f(\boldsymbol{\theta})$ on the sphere can only be sampled at a finite number of discrete directions $\boldsymbol{\theta}_j,\ j = 1, \ldots, J$ by a set of transducers. The integral from Eq. A.14 and the infinite sum from Eq. A.15 turn into finite sums and can therefore be written in matrix vector notation, which yields a linear system of J simultaneous equality constraints [Zot10]

$$\boldsymbol{f} = \mathcal{IDSHT}(\boldsymbol{\gamma}) = \mathbf{Y}_N \boldsymbol{\gamma}_N, \tag{A.16}$$

$$\boldsymbol{\gamma}_N = \mathcal{DSHT}(\boldsymbol{f}) = \mathbf{Y}^+ \boldsymbol{f}, \tag{A.17}$$

with

$$\boldsymbol{f} = \begin{pmatrix} f(\boldsymbol{\theta}_1) \\ f(\boldsymbol{\theta}_2) \\ \vdots \\ f(\boldsymbol{\theta}_J) \end{pmatrix}, \quad \boldsymbol{\gamma}_N = \begin{pmatrix} \gamma_0^0(kr) \\ \gamma{-1}_1(kr) \\ \gamma_1^0(kr) \\ \gamma_1^1(kr) \\ \vdots \\ \gamma_N^N(kr) \end{pmatrix}, \quad \mathbf{Y}_N = \begin{pmatrix} \boldsymbol{y}_N^T(\boldsymbol{\theta}_1) \\ \boldsymbol{y}_N^T(\boldsymbol{\theta}_2) \\ \vdots \\ \boldsymbol{y}_N^T(\boldsymbol{\theta}_J) \end{pmatrix}, \quad \boldsymbol{y}_N(\varphi, \vartheta) = \begin{pmatrix} Y_0^0(\varphi, \vartheta) \\ Y_1^{-1}(\varphi, \vartheta) \\ Y_1^0(\varphi, \vartheta) \\ Y_1^1(\varphi, \vartheta) \\ Y_2^{-2}(\varphi, \vartheta) \\ \vdots \\ Y_N^N(\varphi, \vartheta) \end{pmatrix}. \tag{A.18}$$

For the linear system in Eq. A.16, Eq. A.17 to be solvable the matrix $\mathbf{Y}$ has to be well conditioned so that a stable inverse $\mathbf{Y}^+$ exists and the linear system has to be fully determined or over-determined $J \geq (N+1)^2$. If the assumption of strict band-limitation to order $N$ holds, the spherical spectrum $\boldsymbol{\gamma}_N$ can be calculated exact up to order $N$ [Zot10]. The achievable order $N$ depends on the sampling scheme that is used.

According to Eq. A.16, Eq. A.10 can also be written in matrix vector notation

$$p(kr, \varphi, \vartheta) = \boldsymbol{y}_N^{\mathrm{T}}(\varphi, \vartheta) \underbrace{(\mathrm{diag}\{\boldsymbol{j}_N(kr)\}\boldsymbol{b}_N + \mathrm{diag}\{\boldsymbol{h}_N(kr)\}\boldsymbol{c}_N)}_{\boldsymbol{\psi}_N(kr)}, \tag{A.19}$$

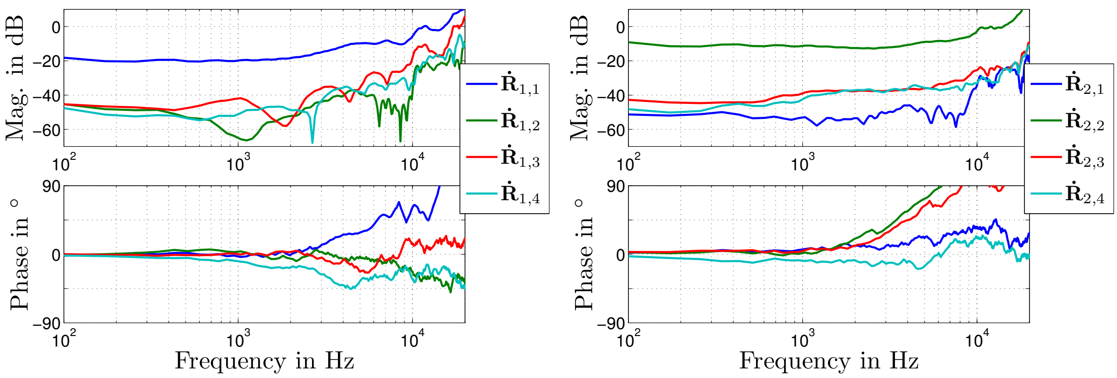
with $\boldsymbol{b}_N = [b_0^0, b_1^{-1}, b_1^0, b_1^1, \ldots, b_N^N]^{\mathrm{T}}$ and $\boldsymbol{j}_N = [j_0, j_1, j_1, j_1, \ldots, j_N]^{\mathrm{T}}$ and writing the sampled pressure as a vector yields

$$\boldsymbol{p} = \boldsymbol{Y}_N \boldsymbol{\psi}_N(kR), \tag{A.20}$$

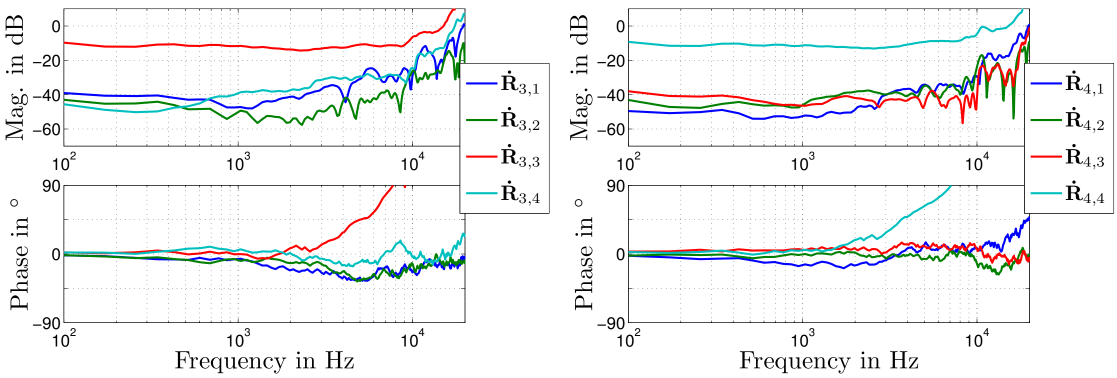
$$\boldsymbol{\psi}_N(kR) = \boldsymbol{Y}_N^{+}\boldsymbol{p}. \tag{A.21}$$

# Appendix B

# Measurement Single Tetrahedral Array

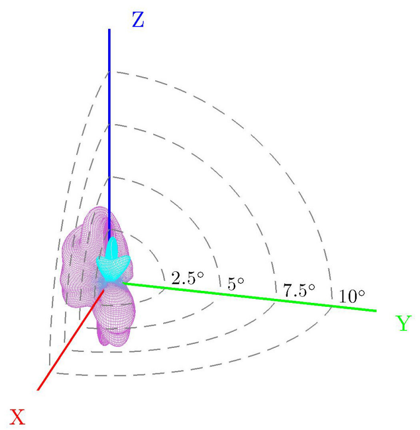



(a) $\dot{\mathbf{R}}_{1,j}$ (b) $\dot{\mathbf{R}}_{2,j}$

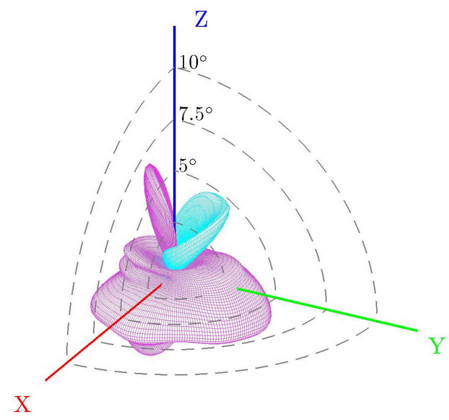



(c) $\dot{\mathbf{R}}_{3,j}$ (d) $\dot{\mathbf{R}}_{4,j}$

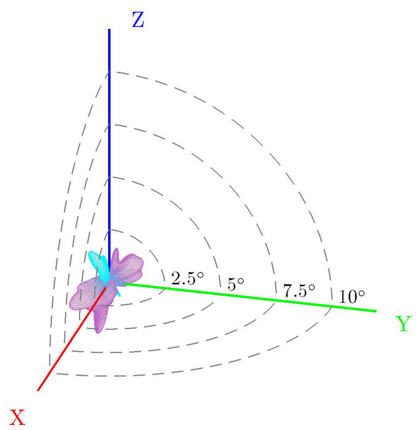
Figure B.1: Frequency response of $i = 1, \ldots, 4$ rows and $j = 1, \ldots, 4$ columns of $(4 \times 4)$ filter matrix $\dot{\mathbf{R}}_{i,j}$ over frequency.



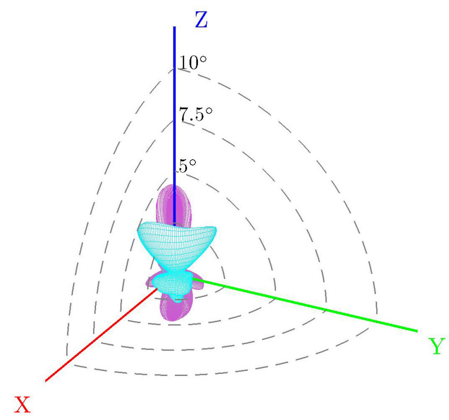

(a) $\Delta\varphi$, based on $\boldsymbol{\chi}_{N_A}(kR)$




(b) $\Delta\vartheta$, based on $\boldsymbol{\chi}_{N_A}(kR)$




(c) $\Delta\varphi$, based on $\hat{\boldsymbol{\phi}}_{N_A}(kR)$

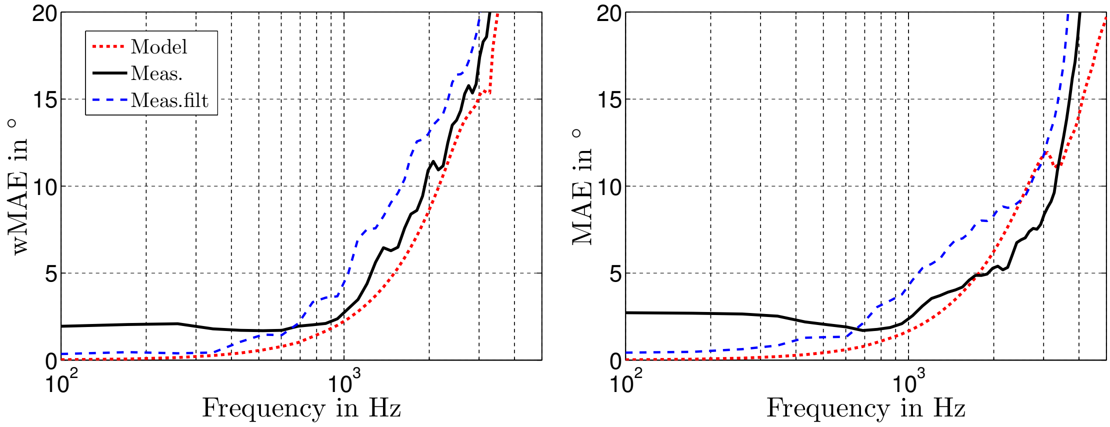



(d) $\Delta\vartheta$, based on $\hat{\boldsymbol{\phi}}_{N_A}(kR)$

Figure B.2: Error pattern of DOA estimation azimuth ($\Delta\varphi$) and zenith error ($\Delta\vartheta$) in °. DOA estimation is done according to Eq. 4.17; (a),(b) based on B-Format signals $\boldsymbol{\chi}_{N_A}$ and (c),(d) based on filtered B-Format signals $\hat{\boldsymbol{\phi}}_{N_A}$ for $f = 350\ Hz$ and plane wave excitation of order $N_C = 10$. (magenta: estimation too high. cyan: estimation too low.)

# B.1 DOA Estimation Example

Let us assume the soundfield of an incident plane wave coming from direction $\boldsymbol{\theta}_i$. The wave spectrum $b_n^m$ of a plane wave is described as $b_n^m = 4\pi i^n\ Y_n^m(\boldsymbol{\theta}_i)$, where $Y_n^m(\boldsymbol{\theta}_i)$ corresponds to the spherical harmonics evaluated at the direction $\boldsymbol{\theta}_i$ of the incident




(a) $\Delta\varphi_{\mathrm{wRMSE}}$ (b) $\Delta\vartheta_{\mathrm{wRMSE}}$

Figure B.3: Weighted MAE of azimuth $\Delta\varphi_{\mathrm{wMAE}}$ and MAE of zenith $\Delta\vartheta_{\mathrm{MAE}}$ DOA estimation error over frequency for plane wave excitation, based on different B-Format signals: Model $\boldsymbol{\chi}_{N_A}$ (Eq. 2.14), Meas.nofilt $\boldsymbol{\chi}_{N_A}$ (Eq. 4.16), Meas.filt $\tilde{\boldsymbol{\chi}}_{N_A}$ (Eq. 4.17). (Array 1)

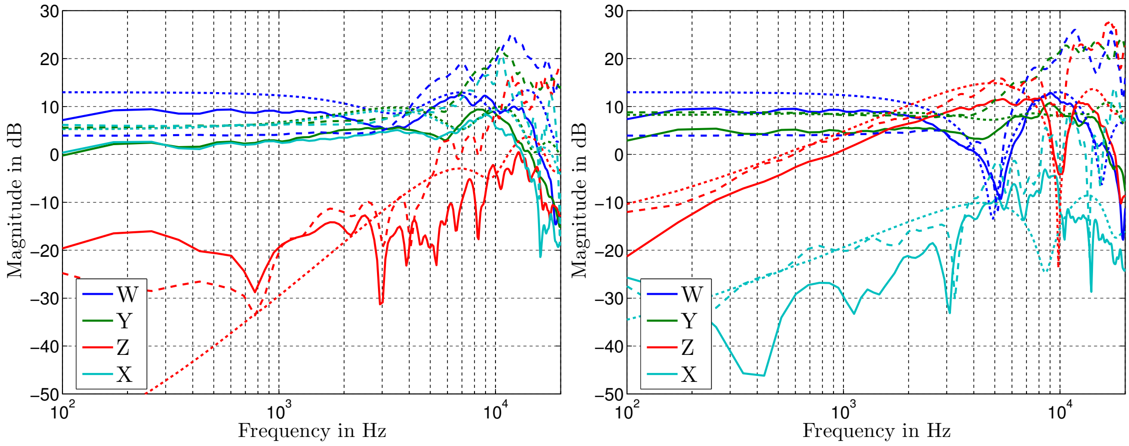
plane wave. In matrix vector notation this yields

$$\boldsymbol{b}_{N_C} = \mathbf{P}_{N_C}\ \mathbf{Y}_{N_C}(\boldsymbol{\theta}_i), \tag{B.1}$$

$$\mathbf{P}_{N_C} = \begin{pmatrix} 4\pi & 0 & \cdots & 0 \\ 0 & 4\pi\mathrm{i}^1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & 4\pi\mathrm{i}^{N_C} \end{pmatrix}. \tag{B.2}$$

Inserting Eq. B.1 into Eq. 2.14 and $\boldsymbol{\phi}_{N_C} = \mathrm{diag}\{\frac{1}{-ik\boldsymbol{h}_{N_C}(kr_0)}\ \boldsymbol{b}_{N_C}$ into Eq. 4.16 we obtain the model and measurement based spherical wave spectrum $\boldsymbol{\chi}_{N_A}$ on the microphone side, where $\boldsymbol{\phi}_{N_C}$ denotes the source distribution at the loudspeaker radius $r_0 = 1.4$ m of a plane wave's wave spectrum $\boldsymbol{b}_{N_C}$ (from Eq. B.1) and $\boldsymbol{h}_{N_C}$ are the spherical Hankel functions of the second kind. Both $\boldsymbol{\chi}_{N_A}$ are compared for two different impinging plane waves with angles of incidence chosen as $(\varphi_i, \vartheta_i) = (45°, 90°)$ and $(\varphi_i, \vartheta_i) = (90°, 90°)$. The magnitudes of the B-Format signals contained in the model and measurement based $\boldsymbol{\chi}_{N_A}$ are shown in Fig. B.4. The figures further illustrate the B-Format signals $\tilde{\boldsymbol{\chi}}_{N_A}$ based on the filtered measurements, which are obtained by inserting $\boldsymbol{\phi}_{N_C}$ into Eq. 4.17. The plane waves were generated with a band limited excitation order of $N_C = 10$. The incident direction $(\varphi_i, \vartheta_i) = (90°, 90°)$

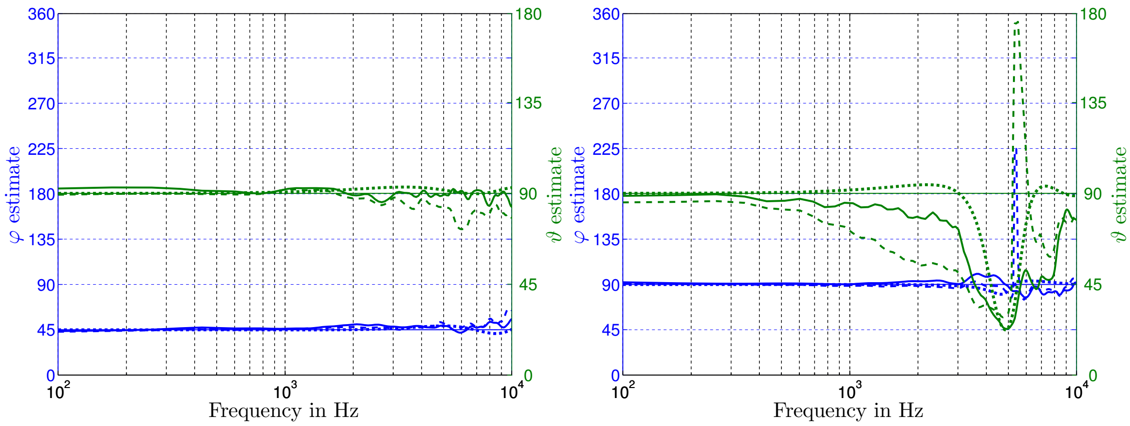
is chosen, because here the effect of spatial aliasing is especially high compared to other directions.[1].




(a) $\boldsymbol{\chi}_{N_A}$ for $(\varphi_i, \vartheta_i) = (45°, 90°)$ (b) $\boldsymbol{\chi}_{N_A}$ for $(\varphi_i, \vartheta_i) = (90°, 90°)$

Figure B.4: B-Format signals $\boldsymbol{\chi}_{N_A} = (\mathrm{W},\ \mathrm{Y},\ \mathrm{Z},\ \mathrm{X})^T$ based on model (dashed line), measurement (solid line) and filtered measurement (dotted line) (see Eq. 4.17) for a plane wave impinging from direction $(\varphi_i, \vartheta_i)$.

For an aliasing free system the Z component should ideally be 0, but due to spatial aliasing it keeps rising with increasing frequency.




(a) $\hat{\varphi}, \hat{\vartheta}$ for $(\varphi_i, \vartheta_i) = (45°, 90°)$ (b) $\hat{\varphi}, \hat{\vartheta}$ for $(\varphi_i, \vartheta_i) = (90°, 90°)$

Figure B.5: DOA estimation angle pairs $\hat{\varphi}, \hat{\vartheta}$ for B-Format signals based measurement (solid line), filtered measurement (dashed line) and model (dotted line) (see Eq. 4.17) for a plane wave impinging from direction $(\varphi_i, \vartheta_i)$.

[1]This can be verified by looking at the directivity pattern of the B-Format's Z component in Fig. 2.8e, which exhibits the highest distortions around the equator at $\varphi_i = 0°, \pm 90°, 180°$

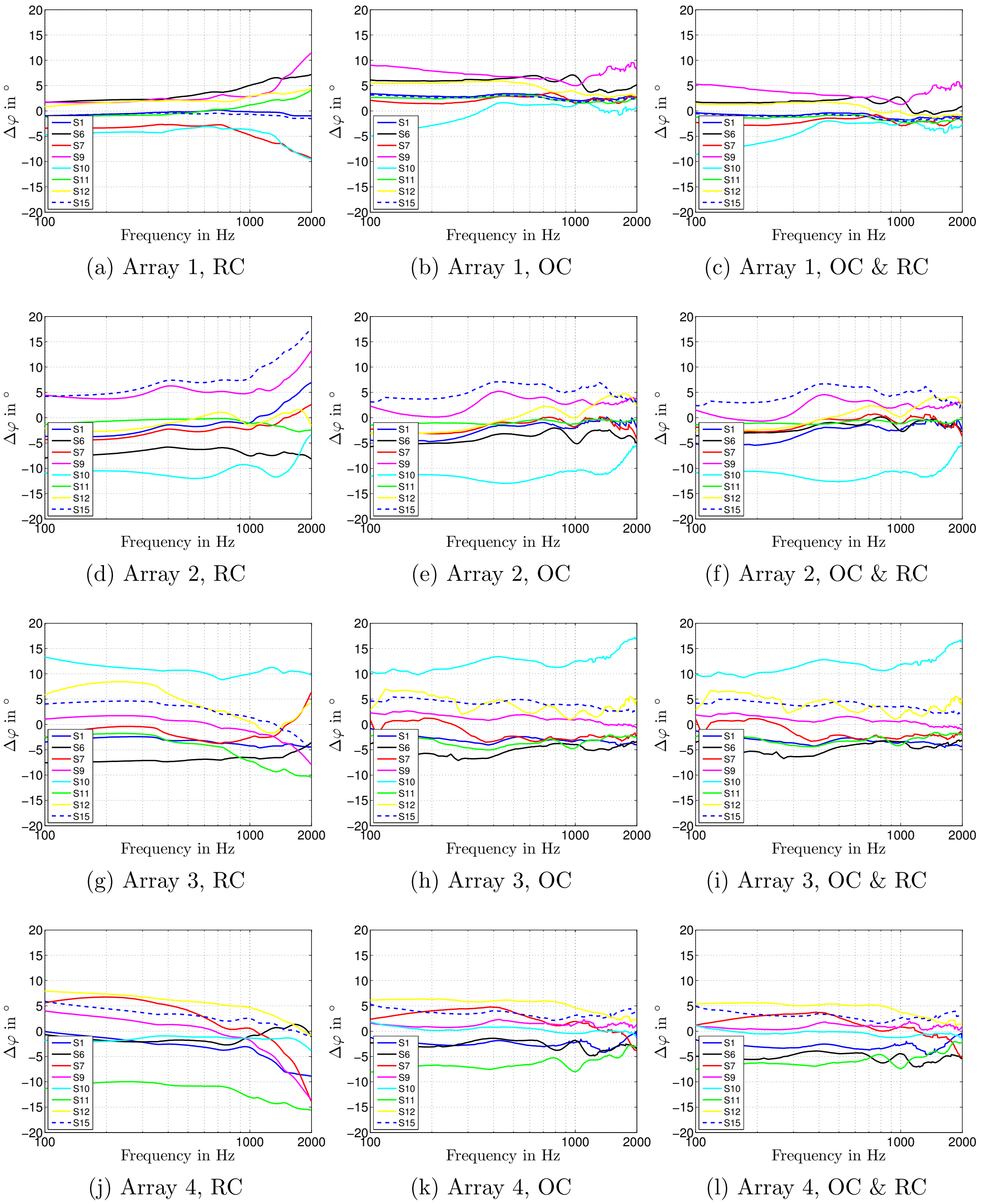
At approximately $500 - 600$ Hz it reaches an amplitude of 15 dB below the W and Y components and starts to influence the DOA estimation illustrated in Fig. B.5. The DOA estimations are based on the intensity vector approach from Eq. 2.1.

# Appendix C

# Experimental Evaluation Figures

| Source | $x$ $[m]$ | $y$ $[m]$ | $z$ $[m]$ |
|---|---|---|---|
| $S_1$ | 0.04 | 0 | 1.51 |
| $S_6$ | 0.04 | 0.5 | 0.91 |
| $S_7$ | -0.5 | -0.47 | 0.91 |
| $S_8$ | -0.57 | -0.49 | 1.58 |
| $S_9$ | 0.83 | -1.77 | 1.48 |
| $S_{10}$ | 0.49 | 1.11 | 1.47 |
| $S_{11}$ | -2.12 | 0.64 | 1.23 |
| $S_{12}$ | -0.96 | 0.37 | 1.36 |
| $S_{13}$ | 2.01 | 0.08 | 1.54 |
| $S_{15}$ | 0.49 | -0.01 | 1.54 |
| $S_{16}$ | 1 | 0 | 1.53 |

Table C.1: Source positions



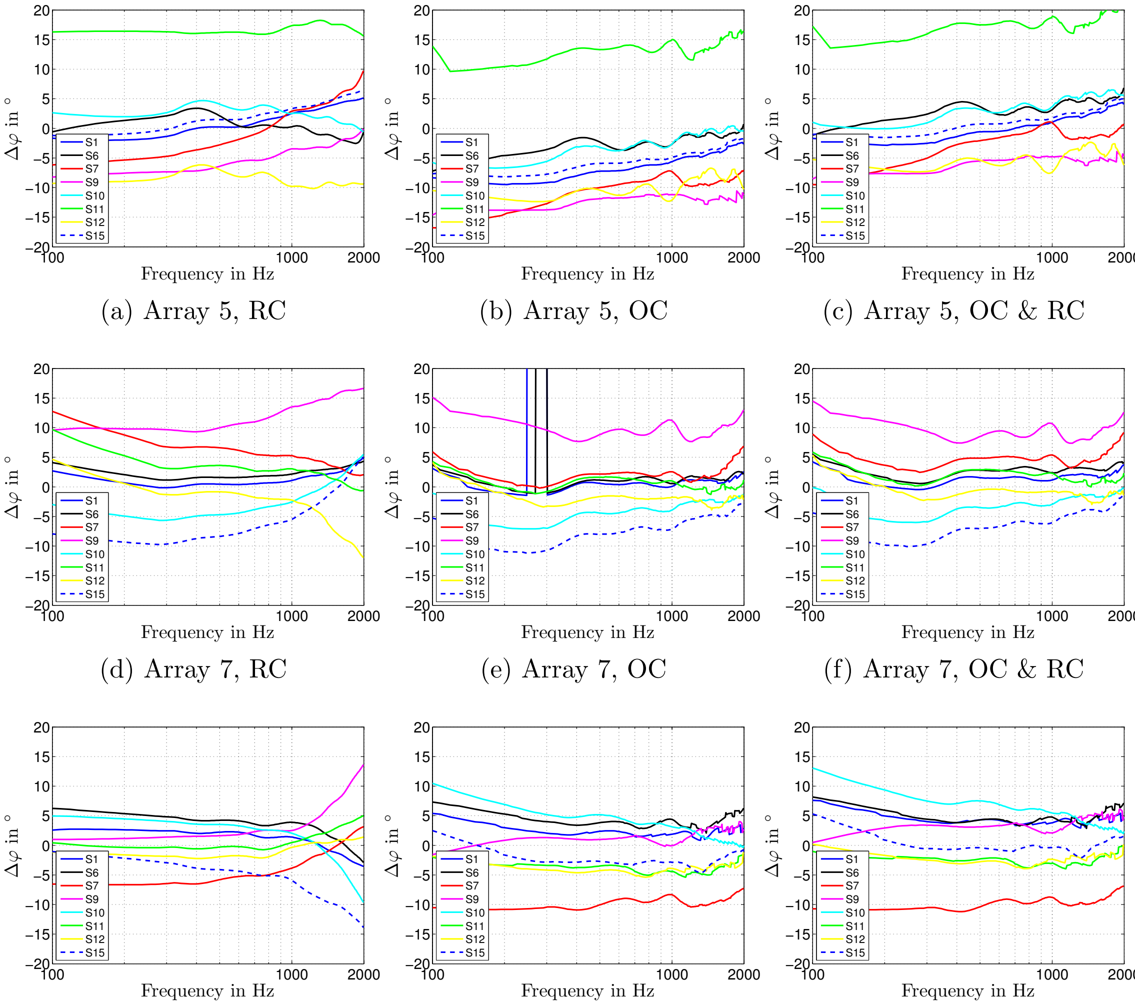

(a) Array 1, RC (b) Array 1, OC (c) Array 1, OC & RC

(d) Array 2, RC (e) Array 2, OC (f) Array 2, OC & RC

(g) Array 3, RC (h) Array 3, OC (i) Array 3, OC & RC

(j) Array 4, RC (k) Array 4, OC (l) Array 4, OC & RC

Figure C.1: DOA azimuth estimation error ($\Delta\varphi$) in $^\circ$ over frequency for microphone arrays $1-4$ for single static sources $S = \{1, 6, 7, 9, 10, 11, 12, 15\}$. The source signals were generated by convolution of a dry male speaker signal and the short (150 Samples) IRs. The illustrated frequency curves denote mean DOA estimations generated from 100 time frames.(a,d,g,j) no correction, (b,e,h,k) with offset correction (OC) and (c,f,i,l) with OC and rotation correction (RC).

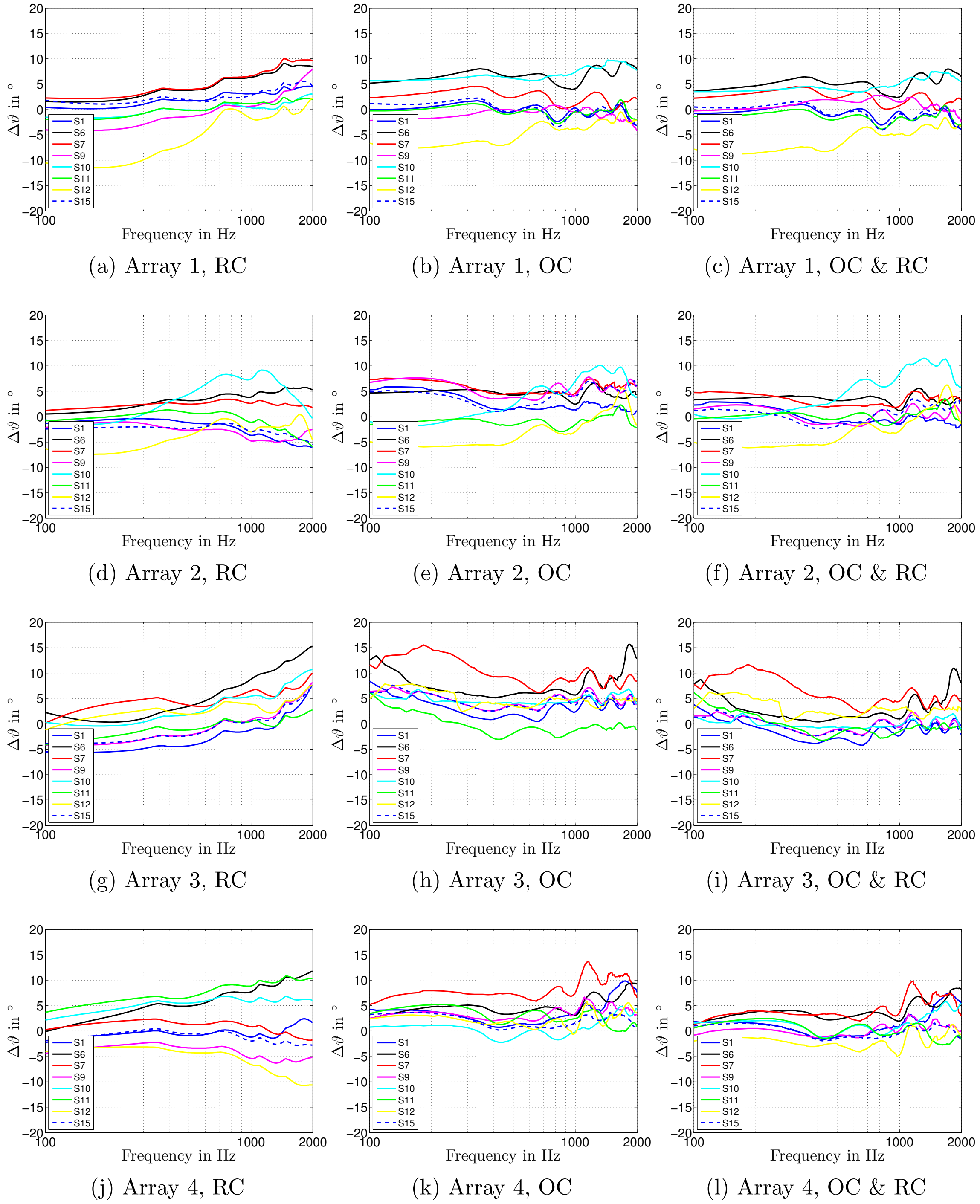



(a) Array 5, RC (b) Array 5, OC (c) Array 5, OC & RC

(d) Array 7, RC (e) Array 7, OC (f) Array 7, OC & RC

(g) Array 8, RC (h) Array 8, OC (i) Array 8, OC & RC

Figure C.2: DOA azimuth estimation error ($\Delta\varphi$) in $^\circ$ over frequency for microphone arrays $5-8$ for single static sources $S = \{1, 6, 7, 9, 10, 11, 12, 15\}$. The source signals were generated by convolution of a dry male speaker signal and the short (150 Samples) IRs. The illustrated frequency curves denote mean DOA estimations generated from 100 time frames.(a,d,g,j) no correction, (b,e,h,k) with offset correction (OC) and (c,f,i,l) with OC and rotation correction (RC).



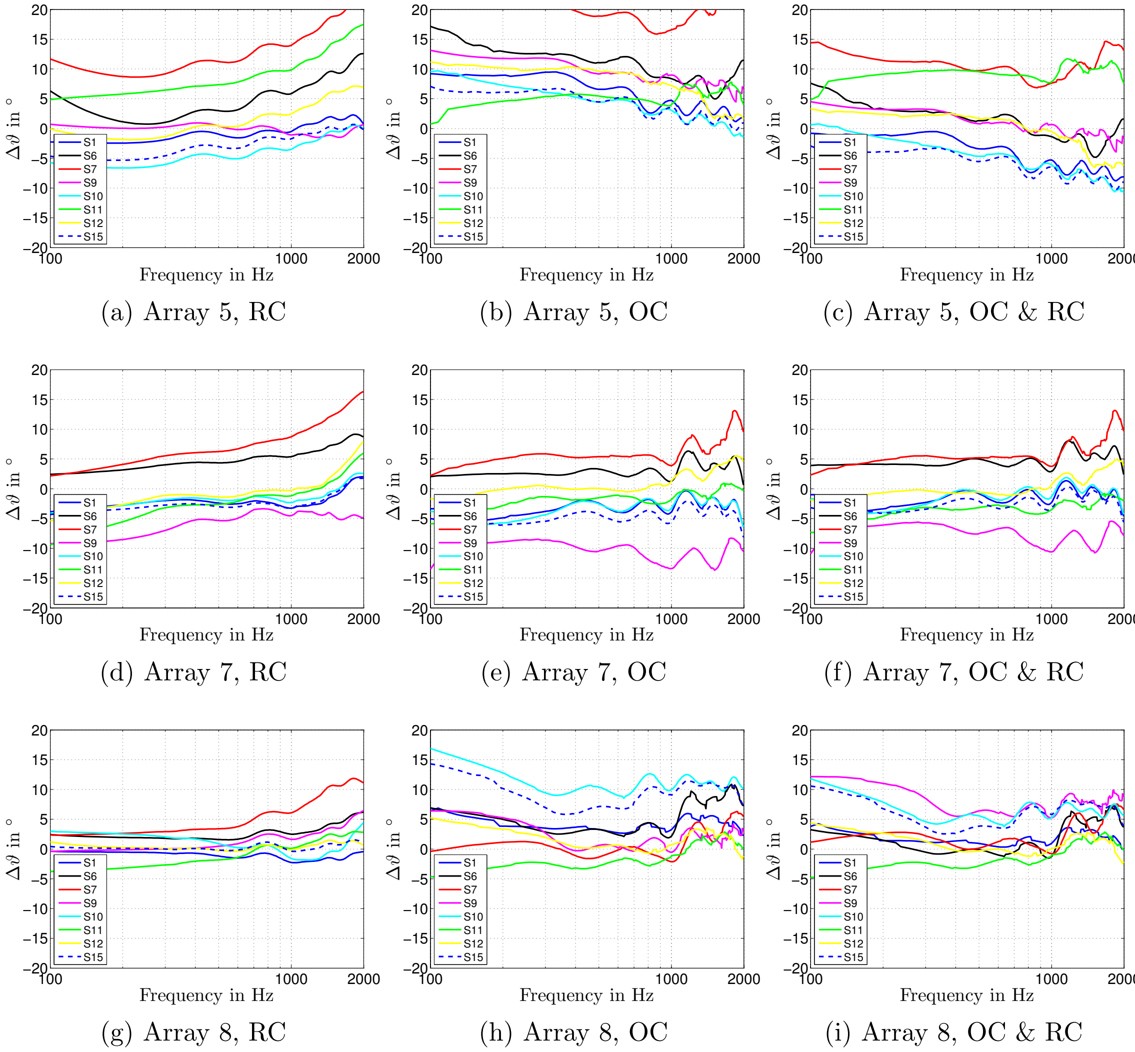

(a) Array 1, RC (b) Array 1, OC (c) Array 1, OC & RC

(d) Array 2, RC (e) Array 2, OC (f) Array 2, OC & RC

(g) Array 3, RC (h) Array 3, OC (i) Array 3, OC & RC

(j) Array 4, RC (k) Array 4, OC (l) Array 4, OC & RC

Figure C.3: DOA zenith estimation error ($\Delta\vartheta$) in $^\circ$ over frequency for microphone arrays $1-4$ for single static sources $S = \{1, 6, 7, 9, 10, 11, 12, 15\}$. The source signals were generated by convolution of a dry male speaker signal and the short (150 Samples) IRs. The illustrated frequency curves denote mean DOA estimations generated from 100 time frames.(a,d,g,j) no correction, (b,e,h,k) with offset correction (OC) and (c,f,i,l) with OC and rotation correction (RC).




(a) Array 5, RC (b) Array 5, OC (c) Array 5, OC & RC

(d) Array 7, RC (e) Array 7, OC (f) Array 7, OC & RC

(g) Array 8, RC (h) Array 8, OC (i) Array 8, OC & RC

Figure C.4: DOA zenith estimation error ($\Delta\vartheta$) in $^\circ$ over frequency for microphone arrays $5-8$ for single static sources $S = \{1, 6, 7, 9, 10, 11, 12, 15\}$. The source signals were generated by convolution of a dry male speaker signal and the short (150 Samples) IRs. The illustrated frequency curves denote mean DOA estimations generated from 100 time frames.(a,d,g,j) no correction, (b,e,h,k) with offset correction (OC) and (c,f,i,l) with OC and rotation correction (RC).

# Bibliography


[APL07] J. Ahonen, V. Pulkki, and T. Lokki, “Teleconference Application and B-format Microphone Array for Direction Audio Coding,” *Audio Engineering Society Conference*, vol. 30, 2007.

[BAS97] M. S. Brandstein, J. E. Adcock, and H. F. Silverman, “A Closed-Form Location Estimator for Use with Room Environment Microphone Arrays,” *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 1, 1997.

[Bat09] J. Batke, “The B-Format Microphone Revised,” *Ambisonics Symposium*, 2009.

[BB07] I. Balmages and R. Boaz, “Open-Sphere Designs for Spherical Microphone Arrays,” *IEEE, Transactions on Audio, Speech and Language Processing*, vol. 15, no. 2, 2007.

[BCH08] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing.* Springer, 2008.

[BH09] J. Batke and H. Hake, “Design Aspects for an Improved B-Format Microphone,” *EUSIPCO - European Signal Processing Conference*, vol. 17, 2009.

[BOS08] A. Brutti, M. Omologo, and P. Svaizer, “Localization of Multiple Speakers Based on a Two Step Acoustic Map Analysis,” *IEEE, Acoustics, Speech and Signal Processing*, pp. 4349–4352, 2008.

[BOS10] ——, “Multiple Source Localization Based on Acoustic Map De-Emphasis,” *EURASIP Journal on Audio, Speech, and Music Processing*, pp. 1–17, 2010.

[BS97] M. S. Brandstein and H. F. Silverman, “A Practical Methodology for Speech Source Localization with Microphone Arrays,” *Computer Speech and Language*, vol. 11, pp. 91–126, 1997.

[BVRB10] C. Bartsch, A. Volgenandt, T. Rohdenburg, and J. Bitzer, "Evaluation of Different Microphone Arrays and Localization Algorithms in the Context of Ambient Assisted Living," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 1–4, 2010.

[cli15] (2015). [Online]. Available: http://cliparts.co/flying-bird-outline

[Far79] K. Farrar, "Soundfield Microphone - Design and Development of Microphone and Control Unit," *Wireless World*, pp. 48–50, 1979.

[Far00] A. Farina, "Simultaneous Measurement of Impulse Response and Distortion With a Swept-Sine Technique," *Audio Engineering Society Convention*, vol. 108, 2000.

[Fre10] K. Freiberger, "Development and Evaluation of Source Localization Algorithms for Coincident Microphone Arrays," Thesis, Institute of Electronic Music and Acoustics (IEM), University of Music and Performing Arts Graz, Technical University Graz, Graz, A, 2010.

[HB04] Y. A. Huang and J. Benesty, *Audio Signal Processing.* Kluwer Academic Publishers, 2004.

[HBE99] Y. A. Huang, J. Benesty, and G. Elko, "Adaptive Eigenvalue Decomposition Algorithm for Real Time Acoustic Source Localization Systems," *IEEE, Acoustics, Speech and Signal Processing*, vol. 2, pp. 937–940, 1999.

[HR08] N. P. Hurley and S. T. Rickard, "Comparing Measures of Sparsity," *Information Theory, IEEE Transactions*, vol. 55, pp. 4723–4741, 2008.

[JHN10] D. Jarrett, E. Habets, and P. Naylor, "3D Source Localization in the Spherical Harmonic Domain Using a Pseudointensity Vector," *Proc. European Signal Processing Conf.(EUSIPCO)*, pp. 442–446, 2010.

[KC76] C. Knapp and G. C. Carter, "The Generalized Correlation Method for Estimation of Time Delay," *Acoustics, Speech and Signal Processing, IEEE*, vol. 24(4), pp. 320–327, 1976.

[Kwo98] Y. Kwon. (1998) Computation of the Rotation Matrix. [Online]. Available: http://www.kwon3d.com/theory/jkinem/rotmat.html

[LD05] Z. Li and R. Duraiswami, “Hemispherical Microphone Arrays for Sound Capture and Beamforming,” *IEEE, Workshop on Applications of Signal Processing to Audio and Acoustics*, 2005.

[MP05] J. Merimaa and V. Pulkki, “Spatial Impulse Response Rendering I: Analysis and Synthesis,” *Journal of the Audio Engineering Society*, vol. 53, no. 12, 2005.

[Pau13] F. Pausch, “A Rigid Double Cone Microphone Array Prototype,” Thesis, Institute of Electronic Music and Acoustics (IEM), University of Music and Performing Arts Graz, Technical University Graz, Graz, A, 2013.

[PF06] V. Pulkki and C. Faller, “Directional Audio Coding: Filterbank and STFT-based Design,” *AES Convention*, vol. 120, 2006.

[PL06] S.-C. Pei and H.-C. Lin, “Minimum-Phase FIR Filter Design Using Real Cepstrum,” *IEEE Transactions on Circuits and Systems*, vol. 53, no. 10, 2006.

[Pom08] H. Pomberger, “Angular and Radial Directivity Control for Spherical Loudspeaker Arrays,” Thesis, Institute of Electronic Music and Acoustics (IEM), University of Music and Performing Arts Graz, Technical University Graz, Graz, A, 2008.

[Raf04] B. Rafaely, “Plane-Wave Decomposition of the Sound Field on a Sphere by Spherical Convolution,” *Journal of the Acoustical Society of America*, vol. 116, 2004.

[Raf05] ——, “Analysis and Design of Spherical Microphone Arrays,” *IEEE on Speech and Audio Processing*, vol. 13, no. 1, 2005.

[SHZ+14] C. Schoerkhuber, P. Hack, M. Zaunschirm, F. Zotter, and A. Sontacchi, “Localization Properties of a New Distributed Array of First-Order Ambisonic Microphones,” *AAAA*, 2014.

[Spa73] D. Sparks, “Algorithm AS 58: Euclidean Cluster Analysis,” *Applied Statistics*, vol. 22, no. 1, pp. 126 – 130, 1973.

[SZZ14] C. Schoerkhuber, M. Zaunschirm, and J. Zmlnig, “WiLMA a Wireless Large-Scale Microphone Array,” 2014.

[TDdS+14] J. Tiete, F. Dominguez, B. da Silva, L. Segers, K. Steenhaut, and T. Abdellah, "Sound Compass: A Distributed MEMS Microphone Array-Based Sensor for Sound Source Localization," *Sensors*, vol. 14, 2014.

[TWH01] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the Number of Clusters in a Data Set via the Gap Statistic," *Journal of the Royal Statistical Society: Series B*, vol. 63, issue 2, pp. 411–423, 2001.

[WC07] H. Wang and P. Chu, "Voice Source Localization for Automatic Camera Pointing System in Videoconferencing," *Audio Engineering Society Conference*, vol. 30, 2007.

[Wil99] E. G. Williams, *Fourier Acoustics.* Academic Press, 1999.

[Zau12] M. Zaunschirm, "Modal Beamforming Using Planar Circular Microphone Arrays," Thesis, Institute of Electronic Music and Acoustics (IEM), University of Music and Performing Arts Graz, Technical University Graz, Graz, A, 2012.

[ZFDZ97] C. Zhang, D. Florencio, E. B. Demba, and Z. Zhang, "Maximum Likelihood Sound Source Localization and Beamforming for Directional Microphone Arrays in Distributed Meetings," *IEEE*, 1997.

[Zot09] F. Zotter, "Analysis and Synthesis of Sound-Radiation with Spherical Arrays," PhD Thesis, University of Music and Performing Arts, Graz, 2009.

[Zot10] ——, "Laborunterlagen Akustische Messtechnik 2, LU, S2010," Ausgabe 2010.

[ZPF09] F. Zotter, H. Pomberger, and M. Frank, "An Alternative Ambisonics Formulation: Modal Source Strenth Matching and the Effect of Spatial Aliasing," *AES Convention Paper*, 2009.

[ZSR03] J. M. Zmoelnig, A. Sontacchi, and W. Ritsch, "The IEM-Cube, a Periphonic Re-/Production System," *AES* $24^{th}$ *International Conference of Multichannel Audio*, 2003.