Elias M. Hoffbauer, BSc
01473027

Evaluation of Surround Sound Setups based on Ambisonic Room Impulse Response Measurements

MASTER'S THESIS

to achieve the university degree of

Diplom-Ingenieur

Inter-university Master's degree programme:

Electrical Engineering and Audio Engineering
(UV 066 413)

at the

University of Music and Performing Arts, Graz
University of Technology, Graz

Supervisor

DI Ph. D. Matthias Frank
Institute of Electronic Music and Acoustics

Assessor

O. Univ. Prof. Mag. art. DI Dr. techn. Robert Höldrich
Institute of Electronic Music and Acoustics

Graz, November 16, 2021

EIDESSTATTLICHE ERKLÄRUNG

Hiermit bestätige ich, dass mir der Leitfaden für schriftliche Arbeiten an der KUG bekannt ist und ich die darin enthaltenen Bestimmungen eingehalten habe. Ich erkläre ehrenwörtlich, dass ich die vorliegende Arbeit selbständig und ohne fremde Hilfe verfasst habe, andere als die angegebenen Quellen nicht verwendet habe und die den benutzten Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Unterschrift

Abstract

In the last decades, spatial audio has become a steadily growing field of research for numerous acoustical disciplines. Following this development, also the fundamental evaluation of loudspeaker setups became more detailed and increasingly complex.
This master's thesis investigates how precise phantom sources can be reproduced on surround sound setups regarding their localization, spatial impression, auditory source width, and coloration. One part of the work is concerned with the direction and position estimation of loudspeakers, which provides the foundation for the prediction of the phantom source localization. In the other part, room acoustics standard measures, such as the lateral energy fraction (LF) and the diffuse-to-reverberant energy ratio (DRR), are adapted for the application at multi-loudspeaker setups. Furthermore, a new measure for the assessment of the listener's envelopment is presented. Based on 1st-order Ambisonic room impulse responses, different spatialization methods are virtually simulated on three differently sized setups and afterwards evaluated with newly developed measures.

Kurzfassung

In den letzten Jahrzehnten wurde die räumliche Wiedergabe von Audiosignalen (Spatial Audio) zu einem stetig wachsenden Forschungsfeld verschiedener akustischer Disziplinen. Besonders die Qualitätsbestimmung von räumlichen Klangreproduktionen ist eine komplexe, multidimensionale Aufgabe, da viele, miteinander verknüpfte Parameter miteinbezogen und ausgewertet werden müssen.
In dieser Masterarbeit wird untersucht, wie präzise Phantomschallquellen hinsichtlich ihres räumlichen Eindrucks, ihrer Quellbreite und Klangfarbe mit Rundumbeschallungsanlagen reproduziert werden können. Ein Teil der Arbeit beschäftigt sich mit der Richtungs- und Positionsschätzung von Lautsprechern, auf deren Basis die Lokalisationsgenauigkeit von Phantomschallquellen beurteilt werden kann. Daran anschließend werden raumakustische Gütemaße, wie z. B. das Seitenschallmaß (LF) und das Direkt-Diffus-Verhältnis (DRR), für die Anwendung bei Surround Sound Systemen mit mehreren Lautsprechern angepasst. Zusätzlich wird ein neues Maß zur Beurteilung des Von-Klang-Eingehüllt-Seins (Listener Envelopment) vorgestellt. Auf der Grundlage von gemessenen Ambisonischen Impulsantworten 1. Ordnung von drei unterschiedlich großen Anlagen werden verschiedene Reproduktionsmethoden virtuell simuliert und anschließend mit den erarbeiteten Maßen evaluiert.

Contents

0. Introduction **1**
- 0.1. A General Approach to Evaluate Surround Sound Setups . . . 1
- 0.2. Structure of the Thesis . . . 2
- 0.3. Terminology and Mathematical Conventions . . . 2

1. Spatial Audio Reproduction **5**
- 1.1. Panning . . . 5
  - 1.1.1. Amplitude Panning Laws . . . 6
  - 1.1.2. Vector-Base Amplitude Panning (VBAP) . . . 7
- 1.2. Ambisonics . . . 9
  - 1.2.1. Spherical Harmonics . . . 9
  - 1.2.2. Common Denominations, Formats and Conventions . . . 10
  - 1.2.3. Ambisonic Decoders . . . 11
    - 1.2.3.1. Sampling Decoder (SAD) . . . 12
    - 1.2.3.2. All-Round Ambisonic Decoding (AllRAD) . . . 13

2. Localization of Sound Events **15**
- 2.1. Brief Note on the Binaural Localization Performance . . . 15
- 2.2. Direction Estimation of Loudspeakers . . . 16
- 2.3. Direction Estimation of Phantom Sources . . . 17
- 2.4. Approaches to a Robust Loudspeaker Localization . . . 19

3. Quality Measures for Surround Sound Setups **21**
- 3.1. Virtual Working Environment . . . 21
- 3.2. Adapted Direct-to-Reverberant Energy Ratio (DRR_m) . . . 23
- 3.3. Adapted Lateral Energy Fraction (LF_m) . . . 28
- 3.4. Measure for Differences in Coloration (C_m) . . . 32
- 3.5. Measure for Envelopment (EV_m) . . . 33
- 3.6. Verification of Measures through Listening Experiments . . . 35

4. Evaluation of Different Surround Sound Setups **37**
- 4.1. Preparation of Measurements . . . 37
- 4.2. Full Evaluation of a Lecture Room: IEM CUBE . . . 39
  - 4.2.1. Localization Accuracy . . . 39
    - 4.2.1.1. Directions of Loudspeakers . . . 40
    - 4.2.1.2. Directions of Phantom Sources . . . 44
    - 4.2.1.3. Conclusion on the Localization Performance . . . 47
  - 4.2.2. Adapted Direct-to-Reverberant Energy Ratio (DRR_m) . . . 48
  - 4.2.3. Adapted Lateral Energy Fraction (LF_m) . . . 51
  - 4.2.4. Differences in Coloration (C_m) . . . 55
  - 4.2.5. Envelopment (EV_m) . . . 57
- 4.3. Exemplary Results for a Concert Hall: György-Ligeti-Hall . . . 59
  - 4.3.1. Localization Accuracy . . . 60
  - 4.3.2. Adapted Lateral Energy Fraction (LF_m) . . . 60

4.3.3. Direct-to-Reverberant Energy Ratio (DRR_m) 63
4.4. Exemplary Results for a Small Studio: Production Studio 64
4.4.1. Localization Accuracy 65
4.4.2. Differences in Coloration (C_m) 66
4.4.3. Envelopment (EV_m) 67

5. Conclusions and Outlook **69**
5.1. A Reasonable Compromise between Simulations and Measurements 69
5.2. Perspectives in Future Research 70

List of Figures **73**

List of Tables **75**

List of Symbols **77**

List of Abbrevations **80**

Bibliography **81**

A. Appendix **87**
A.1. Overview over Ambisonic Formats 87
A.2. Approach of an Improved Position Estimation 88
A.3. Additional Figures for the LF_m Error Estimation 91
A.4. Additional Information to the Evaluations in Chapter 4 92
A.4.1. Additional Figures to the Level Equalization 92
A.4.2. Additional Figures to the Coloration Measure (C_m) 94

0. Introduction

In the last decades, spatial audio has become a steadily growing field of research for scientists of numerous acoustical disciplines, including room acoustics, electroacoustics, as well as psychoacoustics and signal processing. Through qualitative improvements and advancements towards more usability, spatial audio found its place in daily life. Today, multi-loudspeaker systems for spatial audio reproduction are everywhere; they are used in cars, cinemas, concert venues, and in the living rooms of private end users.
Especially the sound reproduction technique *Ambisonics* evolved from a purely academic subject of research to an inherent part of current standards in audio [1, 2] and is used in services and applications of big tech companies such as Google[1] and Facebook[2].

0.1. A General Approach to Evaluate Surround Sound Setups

While the knowledge about spatial audio is steadily extending and improving, the fundamental evaluation of loudspeaker setups becomes increasingly complex.
One approach to ensure a qualitative three-dimensional (3D) sound reproduction is to investigate how accurately phantom sources are spatialized. They should be precisely positioned and their characteristics, such as the apparent source width (ASW), coloration, and sound level, should be as constant as possible. Determining these attributes using room acoustics standard measures implicates time-intense and rather impractical measurement methods. For some measures, such as the perceived source width, even listening experiments can become necessary.
The aim of this work is to use a universal, measure-independent measurement approach in order to simulate and assess the performance of professional and high quality surround sound setups. For the application in this simulated *virtual work environment*, standard room measures, such as the lateral energy fraction (LF), the direct-to-reverberant energy ratio (DRR), and coloration (C), must be adapted. In addition, a localization method for loudspeakers and phantom sources, as well as a new measure for the envelopment of the listeners are presented. On the signal processing side, the focus lies on the spatialization methods *Ambisonics* and *Vector-Base Amplitude Panning* (VBAP).
The proposed approach of evaluation should function as a compromise between pure simulation and a completely hands-on investigation through measurements. On the one hand, the flexibility of a simulation, regarding the retrospective, easy change of parameters, e. g., the used signal processing technique, is ensured. On the other hand, the influence of the room and other intrinsic singularities, which are difficult to simulate, are captured by the measurements.

[1] Google's spatial audio SDK *Resonance*:
https://resonance-audio.github.io/resonance-audio/discover/concepts.html .
[2] Facebook's *360 Spatial Workstation*:
https://facebookincubator.github.io/facebook-360-spatial-workstation/ .

0.2. Structure of the Thesis

Following the presented goals of this work and a note on terminology and mathematical conventions (see below), the remaining content is divided into five chapters.
The first chapter recapitulates the basics of the two spatializiation methods Ambisonics and *Vector-Base Amplitude Panning* (VBAP) on multi-loudspeaker audio systems. Chapter 2 focuses on the localization of loudspeakers and phantom sources. The adaption and development of the four discussed quality measures, Direct-to-reverberant energy ratio (DRR_m), Lateral Energy Fraction (LF_m), Coloration (C_m), and Envelopment (EV_m) are presented in chapter 3. In the subsequent chapter 4, the quality measures are applied to three loudspeaker setups, located in the facilities of the *University of Music and Performing Arts* in Graz: (i) In the lecture room of the *Institute for Electronic Music and Acoustics* (IEM), the so-called *IEM CUBE*, (ii) a concert hall contained in the *House of Music and Music Theater* (MUMUTH) named *György-Ligeti-Hall* and (iii) a small studio facility of the institute, referred to as *Production Studio*. After a short presentation of each space and its technical equipment, the evaluation results are discussed. The localization performance is analyzed in every space and a full assessment of the IEM CUBE, applying all quality measures, is presented. The evaluations of the György-Ligeti-Hall and the Production Studio are focused on two selected measure each in order to proof their adaptability. Following the evaluation, the conclusions of this work and an outlook to further research perspectives are given in the final chapter 5.
In the appendix, an overview of Ambisonic formats is shown and additional figures complement the evaluation in the previous chapters. Furthermore, a theoretical approach for an improved localization method is presented.

0.3. Terminology and Mathematical Conventions

When researching in the field of spatial audio, various systems and nomenclatures can be found in different publications. Hence, this section aims to clarify the used terms, coordinate systems, angles, and notations. The conventions were chosen in order to be as consistent as possible with Zotter's and Frank's publication *Ambisonics, a practical 3D audio theory for recording, studio production, sound reinforcement, and virtual reality* [3]. For a better understanding, a few important terms are explained in more detail in the following.

Discrete and Phantom Sources The term *phantom source* describes sound events that are synthesized between two or more loudspeakers using *amplitude* or *delay panning* [4]. Sound sources which are exhibited by a single loudspeaker at its own position are called *discrete* sound sources in this thesis.

Positions and Directions A clear distinction between the use of *direction* and *position* in this thesis is helpful. Whereas *direction* denotes a vector of unit length $r = 1$ defined by an azimuthal and a zenithal angle in a spherical coordinate system, a point in the three-dimensional space with x-, y-, and z-coordinate of a Cartesian coordinate system is referred to as *position*. Since loudspeakers and microphones have a distinct position in the 3D space, but for Ambisonic encoding and decoding only the two mentioned angles of a direction are required, this differentiation simplifies the description of methods and setups in the following chapters.

Estimated and Measured Locations While assessing the localization performance in chapter 4, the evaluated polar angles and Cartesian coordinates based on impulse response measurements are compared to positions and directions that were conventionally determined by manual distance measuring. In order to preserve clarity in the description, the former evaluated data is denoted as *estimated* directions and positions, whereas the latter is described as *measured* directions and positions.

Impulse Responses and Transfer Functions Ambisonic room impulse responses (ARIRs) are commonly used to describe acoustical parameters of a particular room. In this thesis, the measured impulse responses define primarily the sound propagation from the loudspeakers to the microphone positions. Therefore, they can as well be interpreted as transfer functions, identifying the assessed surround sound system. Both terms are utilized equivalently in here.

Coordinate Systems The coordinate system, used in this work, is defined as follows: The x-axis of the Cartesian and spherical coordinate system points to the front, the y-axis to the left, and the z-axis to the zenith. The azimuthal angle is denoted as $\varphi = [-180°, 180°]$ and $\varphi = 0\,°$ lies on the x-axis. An increase of φ yields a vector turning to the left (anticlockwise). The $0\,°$-direction of the zenith angle $\vartheta = [0°, 180°$ points to the zenith. A zenith angle $\vartheta = 90\,°$ reaches the horizontal plane (see fig. 0.1). The zenith angle can easily be converted into the elevation angle $\tilde{\vartheta}$, which is often used in calculation software, such as MATLAB[3]. $\tilde{\vartheta}$ equals $0\,°$ in the horizontal plane, increases up to $90\,°$ reaching the zenith and decreases to $-90\,°$ while approximating the nadir.

Notation Vectors and matrices are written with bold letters, whereas scalars and functions are defined by thin characters. The alphabetical characters used for matrices are capitalized. Thin capitals stand for the maximum value of a referred variable, e.g., the Ambisonic order N of a signal. For a better readability, the azimuthal and zenithal angle are combined in the direction vector $\boldsymbol{\theta} = [\varphi, \vartheta]$. A Cartesian positional vector $\boldsymbol{\theta}_\mathrm{p} = [x, y, z]$ is indicated by the subscript character p.

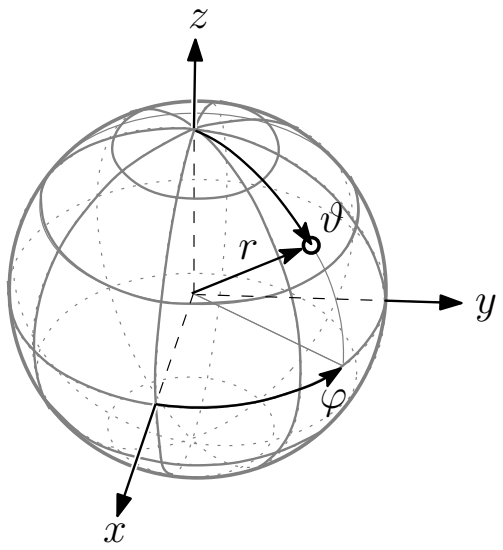


Figure 0.1.: Spherical coordinate system used in this thesis and unit sphere, from [3, p.74]

[3] https://de.mathworks.com/products/matlab.html

1. Spatial Audio Reproduction

There are various approaches to reproduce spatial audio, such as *wave field synthesis* (WFS) [5, 6], *Vector-Base Amplitude Panning*, and *Ambisonics*, as well as proprietary techniques like *Dolby Atmos*[1] or *Vivace* by *Müller-BBM*[2].
The surround systems assessed in this thesis are optimized for the use with VBAP and Ambisonics. Therefore, these two methods for spatializing audio on multi-loudspeaker setups are presented in this chapter. The first section focuses on stereophonic panning and recapitulates on this basis Vector-Base Amplitude Panning by Pulkki [7] briefly. The second part of this chapter serves as a short introduction to the basics of Ambisonics, as well as common systems and methods to decode Ambisonic signals to loudspeakers.

1.1. Panning

Interchannel time differences (ICTD) and *interchannel level differences* (ICLD) between two coherent signals, received at the left and right ear, are crucial measures for determining the direction of a sound source [8, p.140ff].
With the introduction of two-channel sound systems for reproducing sound (*stereophony*) (see fig. 1.1), the possibility of placing sound sources, so-called *phantom sources*, on the tangent between two loudspeakers was established. While phantom sources can be created by inter-loudspeaker time delays and/or level differences, this thesis only focuses on the latter possibility, *amplitude panning* techniques, since *delay panning* is not suitable for large systems that cover a big audience area.

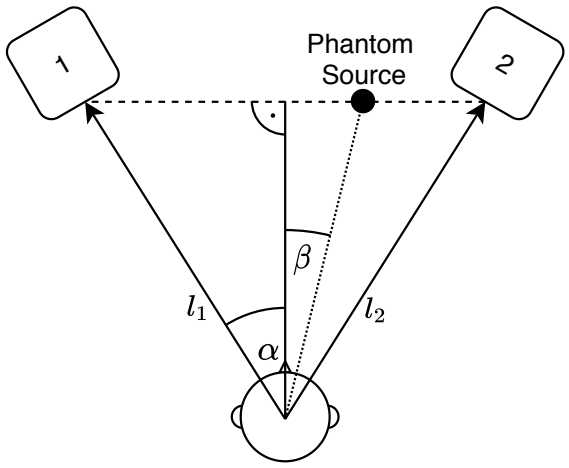


Figure 1.1.: Stereophonic sound system with loudspeaker 1 and 2, aperture angle α and position angle β.

[1] https://www.dolby.com/technologies/dolby-atmos/ .
[2] https://vivace.mbbm-aso.com/de/vivace-de/ .

1.1.1. Amplitude Panning Laws

In order to control the phantom source direction β through adjustable loudspeaker gains g_1 and g_2 on a loudspeaker basis with the aperture angle α (see fig. 1.1), two panning laws were proposed. The Sine Law, derived from simple geometric relations, is defined as [9, 10]:

$$\frac{\sin(\beta)}{\sin(\alpha)} = \frac{g_1 - g_2}{g_1 + g_2}\,. \tag{1.1}$$

In order to incorporate the effect of attenuation around the head, the Sine Law can be modified to the Tangent Law [9]:

$$\frac{\tan(\beta)}{\tan(\alpha)} = \frac{g_1 - g_2}{g_1 + g_2}\,. \tag{1.2}$$

Left and right gain values for an exemplary loudspeaker setup with an aperture angle $\alpha = 45°$ calculated with the Sine and Tangent Law are shown in fig. 1.2.

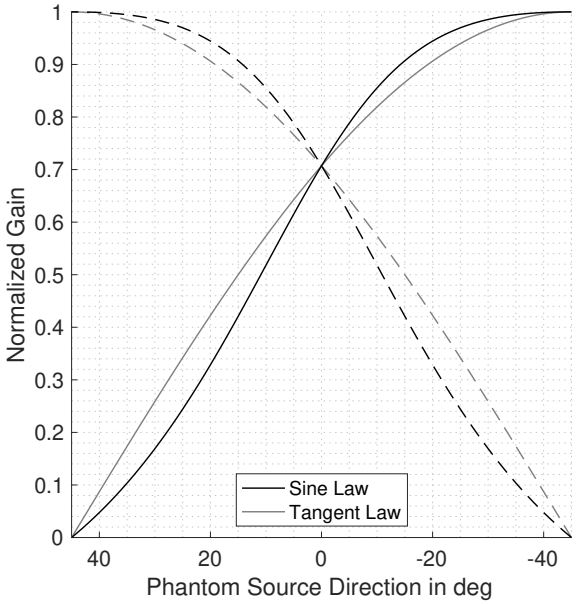


Figure 1.2.: Normalized gains g_1 (dashed line) und g_2 (solid line) of a two-channel sound system (aperture angle $\alpha = 45°$) in dependence of phantom source direction β.

1.1.2. Vector-Base Amplitude Panning (VBAP)

Pulkki reformulated the Tangent Law as a linear combination of two vectors [7]. The vector $\boldsymbol{\theta}_\mathrm{p}$ pointing towards the phantom source, can be expressed by the vectors $\boldsymbol{l}_1$ and $\boldsymbol{l}_2$, which point in the direction of the two loudspeakers and their individual gains g_1 and g_2 [7]:

$$\boldsymbol{\theta}_\mathrm{p} = g_1\boldsymbol{l}_1 + g_2\boldsymbol{l}_2 \ , \tag{1.3}$$

with $\boldsymbol{\theta}_\mathrm{p} = \begin{bmatrix} x_{\boldsymbol{p}} & y_{\boldsymbol{p}} & z_{\boldsymbol{p}} \end{bmatrix}^T$ and $\boldsymbol{l} = \begin{bmatrix} x_{\boldsymbol{l}} & y_{\boldsymbol{l}} & z_{\boldsymbol{l}} \end{bmatrix}^T$.

Panning in a Loudspeaker Triangle With this new mathematical description, the panning law is easily extensible to a set of three loudspeakers $(\boldsymbol{l}_1, \boldsymbol{l}_2, \boldsymbol{l}_3)$. In order to position a phantom source inside a triangle between loudspeakers 1, 2, and 3, its vector $\boldsymbol{\theta}_\mathrm{p}$ is defined by a linear combination of the loudspeaker directions $\boldsymbol{l}_i$ and their gains g_i:

$$\boldsymbol{\theta}_\mathrm{p} = g_1\boldsymbol{l}_1 + g_2\boldsymbol{l}_2 + g_3\boldsymbol{l}_3 \ . \tag{1.4}$$

Written in matrix notation and transformed, the equation yields the gain vector $\boldsymbol{g} = \begin{bmatrix} g_1 & g_2 & g_3 \end{bmatrix}$ consisting of the individual gain values for each loudspeaker.

$$\boldsymbol{g} = \boldsymbol{\theta}_\mathrm{p}^T \boldsymbol{L}_{123}^{-1} \qquad \text{with} \qquad \boldsymbol{L}_{123} = \begin{bmatrix} \boldsymbol{l}_1 & \boldsymbol{l}_2 & \boldsymbol{l}_3 \end{bmatrix}^T \tag{1.5}$$

For most of the phantom source positions, VBAP uses three active loudspeakers ($g_i \neq 0$). However, there are two special cases, where the number of active loudspeakers changes: If the source position $\boldsymbol{\theta}_\mathrm{p}$ equals exactly a loudspeaker position, e. g., $\boldsymbol{l}_1$, just this particular loudspeaker will be active ($g_1 = 1$, $g_2 = g_3 = 0$). If the phantom source is placed on the tangent between two loudspeakers, e. g., 1 and 2, the remaining loudspeaker is inactive ($g_3 = 0$).

Panning on a Multi-Loudspeaker System In this work, the application of VBAP at multi-loudspeaker setups ($L >> 3$) for three-dimensional panning is of interest (see fig. 1.3). On the investigated loudspeaker systems of the IEM CUBE, the György-Ligeti-Hall, and the Production Studio, phantom sources can be positioned on a hemisphere, a so-called *convex hull*, which is defined by the loudspeaker positions of the system and at least one imaginary loudspeaker[3]. Although the number of available loudspeakers is greater than three, a static phantom source is still synthesized using three loudspeakers at most.
To do so, the convex hull is divided into triangular shaped facets[4] with one loudspeaker at each of its three vertices. A loudspeaker can be an element of multiple triangles. Every arbitrary phantom source direction $\boldsymbol{\theta}_\mathrm{p}$ lies now within a specific triangle. According to the individual distance of the phantom source to the three loudspeakers of the specific triangle, the gain vector $\boldsymbol{g}$ is computed with eq. 1.5 . In fig. 1.4, the convex hull and the triangle layout for the loudspeaker arrangement in the György-Ligeti-Hall is shown.

[3] Detailed information to the imaginary loudspeaker is given in sec. 1.2.3.2, *Virtual and Imaginary Loudspeakers*.

[4] In [11], Zotter and Frank recommend the Quickhull Algorithm by Barber et al. [12] for generating the convex hull and grouping loudspeakers into sets of three.

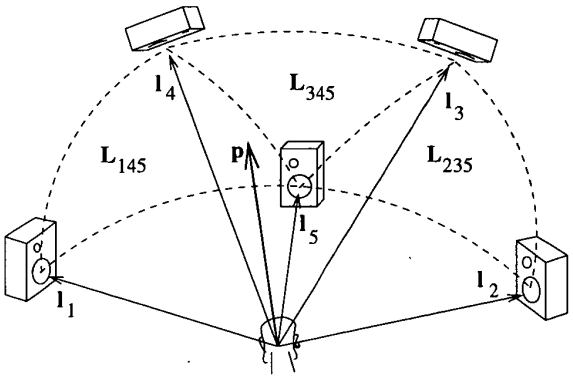


Figure 1.3.: Exemplary Setup for 3D VBAP Using Five Loudspeakers $\boldsymbol{l}_1$-$\boldsymbol{l}_5$, from [7].

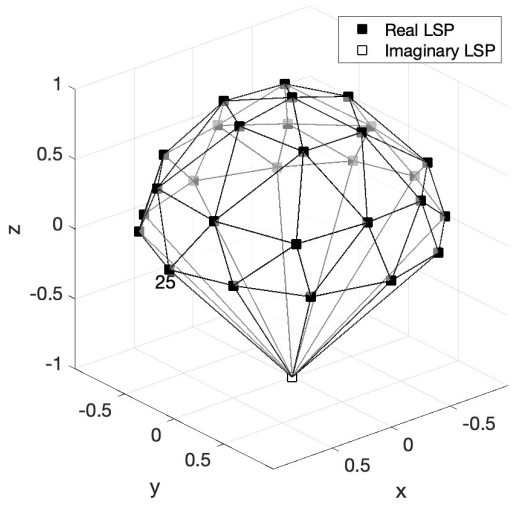


Figure 1.4.: Triangle grouping of loudspeakers (LSPs) in the György-Ligeti-Hall with loudspeaker 25 marking the front.

The panning method VBAP can be implemented as an algorithm as follows: Sequentially, the loudspeaker gains for every triangle based on the source direction $\boldsymbol{\theta}_\mathrm{p}$ are calculated with eq. 1.5. The only triangle which surrounds the phantom source is defined by three loudspeakers with a positive gain value ($g_1 \geq g_2 \geq g_3 \geq 0$). When the all-positive triplet of gains is found, the algorithm stops and returns the three-element gain vector $\boldsymbol{g}$. In order to ensure a constant sound energy, the three gains are normalized so that

$$\sum_{i=1}^{L} g_i^2 = 1 \tag{1.6}$$

is fulfilled.

1.2. Ambisonics

Ambisonics denominates a technique for capturing and reproducing spatial audio while preserving the directional effect and was first introduced by Michal Gerzon in 1975 [13, 14]. The simplest case of Ambisonic recording requires a coincident microphone array with four cardioid capsules oriented in a tetrahedral arrangement [15]. In a second step, the recorded capsule signals can be converted into the Ambisonic domain. The captured characteristics of the sound field are now represented by a set of spherical harmonic functions. For reproducing the recorded surround sound, a suitable decoder is needed that generates the signals for the individual loudspeakers of the Ambisonic reproduction system. While channel-based, discrete surround sound techniques are rather limited by their recording setup to a specific playback system, the Ambisonic format offers more flexibility. Once the recordings are encoded in the Ambisonic domain, they can be reproduced binaurally on headphones or on any arbitrary loudspeaker setup.
For reproducing sound fields in one plane with no height information (*pantophonic*), a 2D implementation of Ambisonics can be used. In this thesis, the term Ambisonics refers only to the 3D application of the technique on a full sphere (*periphonic*).
The subsequent section 1.2.1 describes the mathematical basics of Ambisonics.
Section 1.2.2 presents briefly the most important denominations, formats and conventions in the context of Ambisonics. Decoding loudspeaker signals from Ambisonic signals is the topic of the last section 1.2.3.

1.2.1. Spherical Harmonics

Spherical harmonics are used in several cases for the definition of field properties in scientific disciplines, e. g., for the description of electrostatic fields in physics or magnetic fields in geodesy. In acoustics, spherical harmonics are solutions in the azimuthal and zenithal dimension for the Laplace operator in the Helmholtz equation formulated in spherical coordinates. In that context, they describe the pressure distribution of a sound field on a spherical surface. The interested reader can find an extensive derivation in [3, p.186ff].
The spherical harmonics Y_n^m of order n and degree m can be separated in an azimuthal part $\Phi_m(\varphi)$ and a zenithal part $\Theta_n^m \cos(\vartheta)$:

$$Y_n^m(\varphi, \vartheta) = \Phi_m(\varphi)\Theta_n^m \cos(\vartheta)\,. \tag{1.7}$$

$\Phi_m(\varphi)$ consists of circular harmonics only dependent of the azimuthal angle φ. The zenith angle dependent part Θ_n^m is defined by associated Legendre polynomials P_n^m, which are additionally normalized by the term N_n^m. Therefore, the general definition of Y_n^m equals

$$Y_n^m(\varphi, \vartheta) = \Phi_m(\varphi)N_n^{|m|}P_n^{|m|} \cos(\vartheta)\,. \tag{1.8}$$

Fig. 1.5 represents the spherical harmonics from the 0th to the 5th order. The number of degrees m increases with the order n and ranges over the interval $m = [-n, n]$.
With a sum of spherical harmonics of infinite order, every sound field can be represented ideally and without loss of information, under the condition that the directional functions γ_{nm} representing the sound field are square-integrable [3, p.74f].

$$g(\varphi, \vartheta) = \sum_{n=1}^{\infty} \sum_{m=-n}^{n} \gamma_{nm} Y_n^m(\varphi, \vartheta) \tag{1.9}$$

Obviously, only a finite number of orders is available and therefore just an approximation of the original sound field can be reproduced in practical applications. Nevertheless, good results can already be achieved with a reasonable high order of spherical harmonics. (see sec. 1.2.2, *Higher-Order Ambisonics (HOA)).*

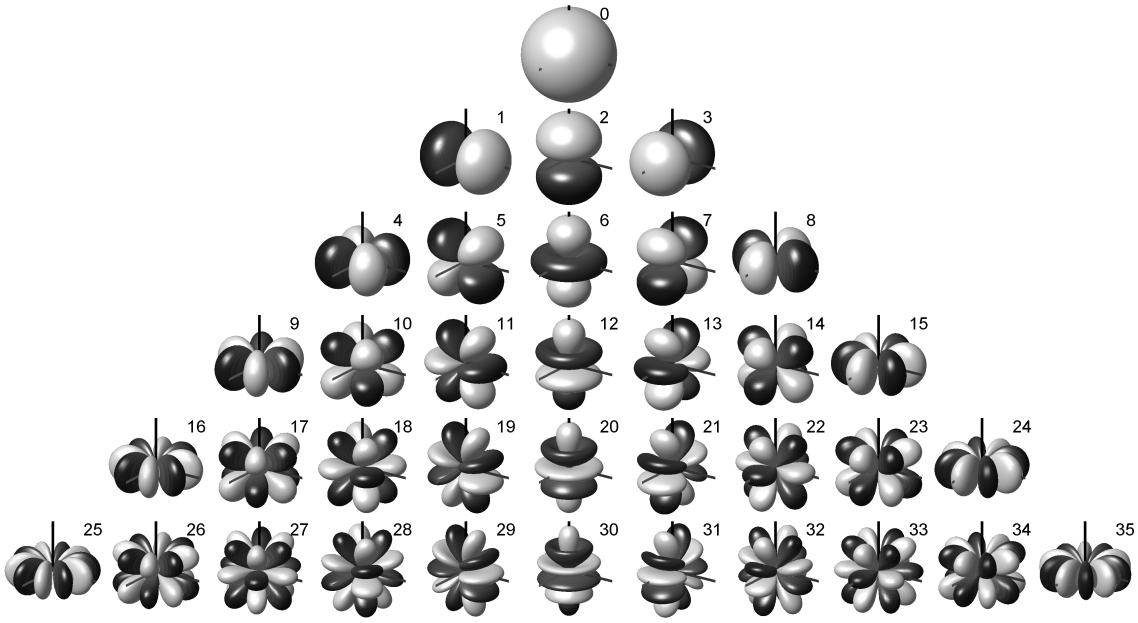


Figure 1.5.: Spherical harmonics of the Ambisonic orders 0 to 5 with positive (bright surface) and negative sign (dark surface), from [16].

1.2.2. Common Denominations, Formats and Conventions

First-Order Ambisonics (FOA) is the lowest order that can preserve directional features of the recorded sound field. It is common use to name its four channels after the four last letters of the alphabet W, X, Y, and Z. The W-channel holds the omnidirectional signal of the 0th order, while the three remaining channels of the 1st order comprise the signals recorded with a figure-of-eight pickup pattern oriented in the three spatial directions x, y, and z. A 1st-order recording can be achieved by using for example four cardioid microphones arranged in a tetrahedral microphone array [3, p.11]. There are also different compact solutions available, such as the Rode NT-SF1[5] or the Soundfield ST450 MKII[6], which was used for the measurements evaluated in this work.

Channel Ordering During the development of Ambisonics, multiple ways of sequencing the channels were proposed. In the traditional *B-Format*, the channels are ordered after the alphabet ($W(t)$, $X(t)$, $Y(t)$, and $Z(t)$). With the extension of the Ambisonics technique to higher orders, this naming convention became increasingly impractical due to the more than quadratically rising channel number ($N_{\mathrm{ch}} = (N+1)^2$). For higher orders, the system of the *Ambisonic Channel Number* (ACN) proves to be easily extensible. Every channel is assigned to a unique number, which is calculated with the degree n and rank m of its related spherical harmonic Y_n^m [17]:

$$\mathrm{ACN} = n(n+1) + m \ , \ \mathrm{ACN} \in \mathbb{N} \ . \tag{1.10}$$

Normalization Regarding 1st-order Ambisonics, Gerzon proposed that the W-channel should be damped by the factor $\frac{1}{\sqrt{2}}$ [14], which was adapted by Malham and Furse in their *FuMa*-weighting-convention and later extended up to the third Ambisonics order [18]. Daniel presents in his work two further variants [19]: The *full 3D-normalization*

[5] https://www.rode.com/microphones/ntsf1 .
[6] https://www.soundfield.com/#/products/st450mk2 .

(N3D)

$$\boldsymbol{a}_{\mathrm{N3D}}(n,m) = \sqrt{\frac{(2-\delta_m)(2n+1)}{4\pi}\frac{(n-|m|)!}{(n+|m|)!}} \tag{1.11}$$

and the *Schmidt semi-normalization* (SN3D):

$$\boldsymbol{a}_{\mathrm{SN3D}}(n,m) = \sqrt{\frac{(2-\delta_m)}{4\pi}\frac{(n-|m|)!}{(n+|m|)!}} = \boldsymbol{a}_{\mathrm{N3D}}\frac{1}{\sqrt{2n+1}} \tag{1.12}$$

with the Kronecker delta $\delta_{m0} = 1$. The full normalization N3D yields an equal power distribution in all channels, if one assumes the described sound field as a diffuse field [19, p.155 ff.]. These two normalizations are easily convertible using the factor $\frac{1}{\sqrt{2n+1}}$. The factor 4π is derived from the surface of a full unit sphere (radius $r = 1$) and is adapted from Zotter and Frank [3, p.130] and Nachbar [20].

Formats Two formats for Ambisonic signals are the most widespread: For lower orders, the B-format is often used[7]. In B-formatted files, the channels are assigned with letters, ordered alphabetically and weighted in the FuMa-style. Ambisonic scenes are typically not directly recorded in the B-format. The individual B-format-channels W, X, Y and Z can be computed through a linear combination of the recorded capsule signals, called *A-format*.
If not otherwise declared, the ambiX-format, proposed by Nachbar et al., is used in this thesis [20], which means that channels are ordered according to ACN and normalized with SN3D gains. In appendix A.1, table A.1 gives a detailed overview over the two formats.

Higher-Order Ambisonics (HOA) describes all orders $N \geq 2$ and has several advantages compared to 1st-order Ambisonics: It allows to represent sound fields with a sharper directional resolution, clearer impression of depth and also the sweet spot area of loudspeaker setups increases with the Ambisonic order [21, 22]. A disadvantage is the fast increasing amount of data, since the channel number increases order dependent with $(N+1)^2$. With the disappearing limits of storage capabilities and processing power of current computers, a lot of research went into utilizing those benefits in the last decades. HOA microphones such as the em32 Eigenmike[8] or the Zylia ZM-1[9] enable spatial audio recordings in higher-order Ambisonic. Furthermore, there are numerous upmix algorithms that try to enhance 1st-order signals or measured Ambisonic room impulse responses (ARIRs) to higher-order spherical harmonics [23, 24, 25, 26].

1.2.3. Ambisonic Decoders

For reproducing Ambisonic signals on loudspeaker arrangements, an individual decoder has to be designed so that it fits the playback system optimally, since it is dependent on the specific directions of the loudspeakers. Mathematically, the decoding process itself equals a multiplication of a so-called decoding matrix and the Ambisonic signal.
Besides numerous decoding approaches [27, 3, p. 78ff], e. g., *Mode-Matching* (MAD) [19, 28, 29] or *Energy-Preserving Ambsionic Decoding* (EPAD) [30], this section presents the basic *Sampling Decoder* (SAD) and the *All-Round Ambisonic Decoding* (AllRAD), which was mainly used in this thesis.

[7] For example, the preamplifier of the Soundfield ST450 MKII microphone provides Ambisonic scenes in B-format.
[8] https://mhacoustics.com/products .
[9] https://www.zylia.co/zylia-zm-1-microphone.html .

1.2.3.1. Sampling Decoder (SAD)

Using the Sampling Decoder is the simplest way to decode an Ambisonic signal $\boldsymbol{\chi}$ to an arbitrary number L of loudspeakers placed in the directions $\boldsymbol{\theta}_l$.
The weighting matrix $\boldsymbol{Y}_N$ is constructed by evaluating the spherical harmonics Y_n^m for the direction $\boldsymbol{\theta}_l(\varphi_l, \vartheta_l)$ of every loudspeaker l:

$$\boldsymbol{Y}_N = [\boldsymbol{y}_N(\boldsymbol{\theta}_1), \ldots, \boldsymbol{y}_N(\boldsymbol{\theta}_L] \, . \tag{1.13}$$

By inserting the vector $\boldsymbol{y}_N(\boldsymbol{\theta}_l) = [Y_0^0(\boldsymbol{\theta}_l), Y_1^{-1}(\boldsymbol{\theta}_l), \ldots, Y_N^N(\boldsymbol{\theta}_l)]^T$, one gets the full form of

$$\boldsymbol{Y}_N = \begin{bmatrix} Y_0^0(\boldsymbol{\theta}_1) & Y_0^0(\boldsymbol{\theta}_2) & \ldots & Y_0^0(\boldsymbol{\theta}_L) \\ Y_1^{-1}(\boldsymbol{\theta}_1) & Y_1^{-1}(\boldsymbol{\theta}_2) & \ldots & Y_1^{-1}(\boldsymbol{\theta}_L) \\ \vdots & \vdots & & \vdots \\ Y_N^N(\boldsymbol{\theta}_1) & Y_N^N(\boldsymbol{\theta}_2) & \ldots & Y_N^N(\boldsymbol{\theta}_L) \end{bmatrix} \, . \tag{1.14}$$

For the preservation of the signal energy, a correction factor is added, and the complete decoder matrix $\boldsymbol{D}$ can be formulated as:

$$\boldsymbol{D} = \sqrt{\frac{2}{L}} \boldsymbol{Y}_N^T \, . \tag{1.15}$$

Subsequently, the loudspeaker signals $\boldsymbol{s}_L$ are computed by multiplying the decoder matrix $\boldsymbol{D}$ with the Ambisonic signal $\boldsymbol{\chi}_N$:

$$\boldsymbol{s}_L = \boldsymbol{D} \boldsymbol{\chi}_N \, . \tag{1.16}$$

Weighting By applying weights on the spherical harmonics in the decoding process, different directivity patterns into the direction that is decoded can be achieved. If no individual weights ($\boldsymbol{a}_N = 1$) are applied, the term *basic* weighting is used.
The so-called *max-$\boldsymbol{r}_E$*-weighting produces a beam pattern that maximizes the length of the $\boldsymbol{r}_\mathrm{E}$-vector and suppresses strongly side lobes:

$$\boldsymbol{a}_{N,\text{max-}\boldsymbol{r}_\mathrm{E}}(n,m) = P_n \left[\cos\left(\frac{137.9^\circ}{N + 1.51} \right) \right] \, . \tag{1.17}$$

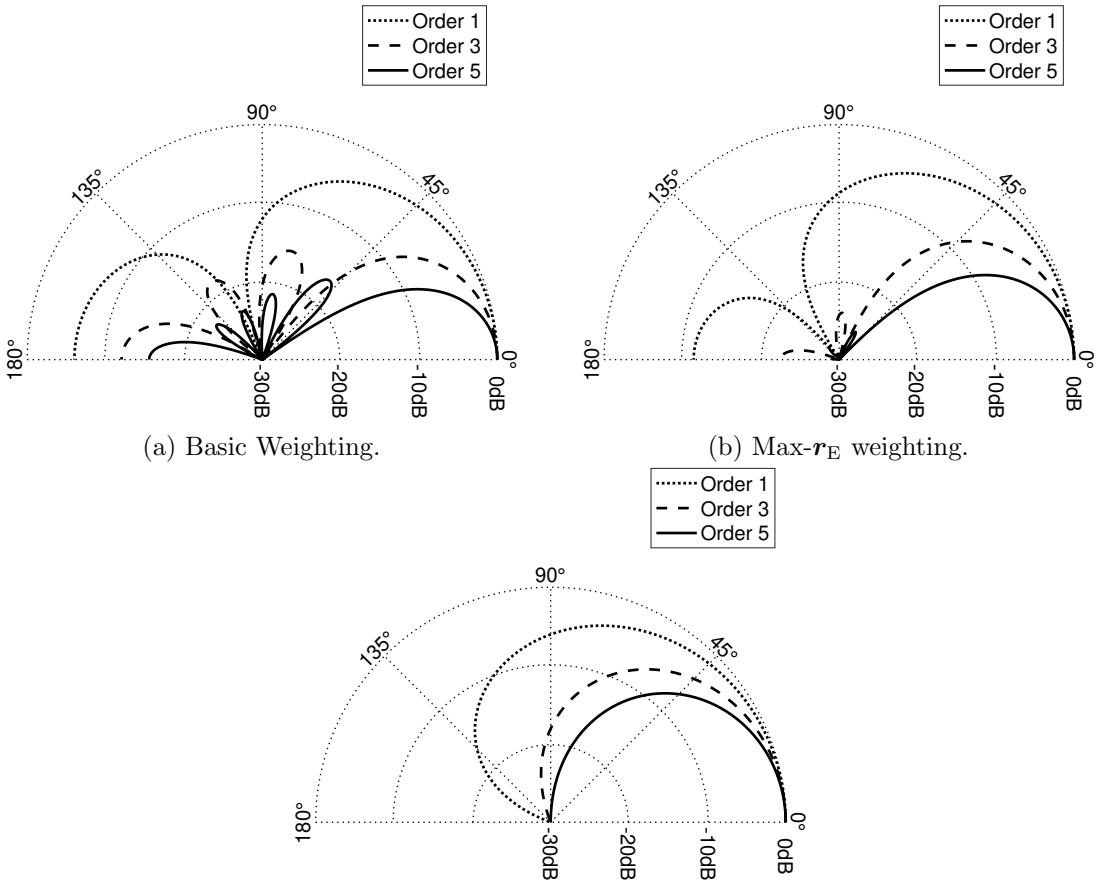
This can be very helpful in practical applications: For example, the max-$\boldsymbol{r}_\mathrm{E}$ weighting can effectively improve the localization of phantom sources at off-center listening positions [3, p.70, 31].
The *in-phase* weighting generates a directivity pattern with no side-lobes at the cost of a very wide main lobe (see fig. 1.6c). The order and degree dependent weights can be calculated with eq. 1.18 [19, p.182]:

$$\boldsymbol{a}_{N,\text{in-phase}}(n,m) = \frac{N!(N+1)!}{(N+n+1)(N-m)!} \, . \tag{1.18}$$

In order to apply these patterns directly in the decoding formula, eq. 1.16 can be extended by the weights $\boldsymbol{a}_N$ in diagonal matrix form:

$$\boldsymbol{s}_L = \boldsymbol{D} \, \mathrm{diag}\{\boldsymbol{a}_N\} \boldsymbol{\chi}_N \tag{1.19}$$

The following figures 1.6a, 1.6b and 1.6c present the different symmetrical directivities at the Ambisonic orders N =1, 3, and 5.



(a) Basic Weighting.

(b) Max-$\boldsymbol{r}_{\mathrm{E}}$ weighting.

(c) In-phase Weighting

Figure 1.6.: Directivity Patterns for different weightings at Ambisonic orders 1, 3 and 5.

1.2.3.2. All-Round Ambisonic Decoding (AllRAD)

The AllRAD decoder designed by Zotter and Frank [11] decodes Ambisonic scenes on any arbitrary loudspeaker arrangement using Vector-Base Amplitude Panning by Pulkki [23]. The decoding process consists of two stages: First, the Ambisonic signal is decoded by use of the SAD on a dense set of J virtual loudspeakers

$$\boldsymbol{s}_J = \hat{\boldsymbol{Y}}_N^T \boldsymbol{\chi}_N = [\boldsymbol{y}_N(\hat{\boldsymbol{\theta}}_{j=1}), \ldots, \boldsymbol{y}_N(\hat{\boldsymbol{\theta}}_J)]\boldsymbol{\chi}_N \,. \tag{1.20}$$

The signals $\boldsymbol{s}_J$ of the virtual loudspeakers are then distributed on the L real loudspeakers via VBAP (see sec. 1.1.2). This is done with the help of a so-called VBAP rendering matrix $\hat{\boldsymbol{G}}$ of dimension $L \times J$. The matrix assigns the gains $\boldsymbol{g}_j = \begin{bmatrix} g_{j,1} & \ldots & g_{j,L} \end{bmatrix}^T$ of the physical loudspeakers to its belonging virtual loudspeakers j. Therefore, the signals of the real loudspeakers can be computed as

$$\boldsymbol{s}_L = \begin{bmatrix} \boldsymbol{g}_{j=1} & \ldots & \boldsymbol{g}_J \end{bmatrix} \boldsymbol{s}_J = \hat{\boldsymbol{G}} \boldsymbol{s}_J = \hat{\boldsymbol{G}} \hat{\boldsymbol{Y}}_N^T \boldsymbol{\chi}_N \,. \tag{1.21}$$

Eq. 1.21 has the form of a conventional decoding matrix (see eq. 1.16). Comparing the coefficients of these two equations and adding an energy-preserving factor gives the decoding matrix of the AllRAD decoder

$$\boldsymbol{D} = \frac{2}{L} \hat{\boldsymbol{G}} \hat{\boldsymbol{Y}}_N^T \,. \tag{1.22}$$

The *AllRADecoder* plug-in of the *IEM Plug-in Suite*[10] uses a 5200-point t-design[11] with $t = 100$, published by Gräf and Potts [32]. In order to produce consistent decoding results in MATLAB, as well as in the DAW *Reaper*[12], the decoding matrix of the AllRADecoder plug-in was used for further investigations and simulations in this thesis. Fig. 1.7 and 1.8 show the distribution of the virtual loudspeakers in accordance to the 5200-point t-design and the directions of the 24 loudspeakers in the IEM CUBE and one imaginary loudspeaker mapped onto the unit sphere.

Virtual and Imaginary Loudspeakers Although both types of loudspeakers do not have a counterpart in the physical playback system, it is important to distinguish between *virtual* and *imaginary* loudspeakers in the context of Ambisonic decoders.
AllRAD decodes signals for the virtual loudspeakers by using their positions, which are then panned on the triangle that surrounds the particular virtual loudspeaker. Problems with asymmetrically designed playback setups are thereby eliminated, since virtual loudspeakers can be always ideally positioned.
The addition of imaginary loudspeakers in the nadir to a hemispherical loudspeaker arrangement (see fig. 1.8) helps to define the convex hull unambiguously. Furthermore, phantom sources outside of the panning range (e. g., below the horizontal plane in case of a hemispherical setup), can be reproduced, if a localization mismatch is accepted [11]. Romanov et al. use the insertion of imaginary loudspeakers for VBAP and Ambisonic AllRAD decoding on surround and surround-with-height loudspeaker configurations defined in Rec. ITU-R BS.2051 [33]. Thus, instabilities of the phantom source attributes, including localization, width, and coloration, caused by asymmetries of the setup or the computation of the convex hull can be eliminated [34]. In contrast to signals of virtual loudspeakers, the signal of the imaginary loudspeaker is of no further benefit and therefore not even decoded.

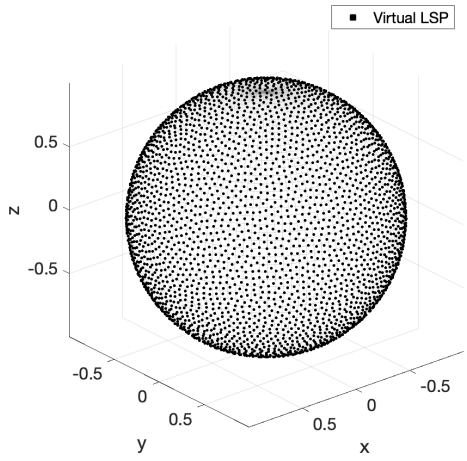


Figure 1.7.: Distribution of 5200 virtual loudspeakers according to the 5200-Point 100-design.

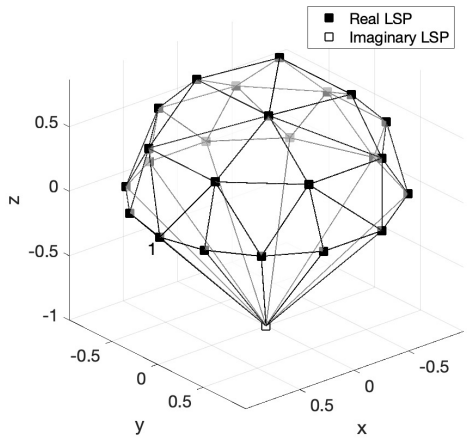


Figure 1.8.: Convex hull and loudspeakers triangles of the system in the IEM CUBE.

[10] https://plugins.iem.at/docs/allradecoder/
[11] The exact coordinates of the 100-design and several others can be found on M. Gräf's website: https://www-user.tu-chemnitz.de/~potts/workgroup/graef/computations/pointsS2.php .
[12] https://www.reaper.fm/

2. Localization of Sound Events

The overall topic of this chapter is the localization loudspeakers and phantom sources. Following a brief excursus on the localization performance of listeners, an approach for determining the direction of loudspeakers using a single microphone array position is introduced.
In section 2.3, findings in literature to the extended $\boldsymbol{r}_\mathrm{E}$-vector model, which uses the estimated loudspeaker directions in order to predict the localization of phantom sources on surround sound systems, are presented. At the end of this chapter, recommendations for a robust loudspeaker localization are summarized (see sec. 2.4).

2.1. Brief Note on the Binaural Localization Performance

The binaural localization of sound sources in the horizontal and vertical dimension works differently. Whereas the azimuthal shift of an auditory event is perceived by *interaural time differences* (ITD) and *interaural level differences* (ILD) [8], the elevation of sound sources is determined by spectral cues produced by the reflections of the pinna, head and torso [35, 36, 37]; however, it is not entirely clear, which specific cues are used exactly for the detection [38].
Listening tests with horizontal, inclined, and vertical stereophonic pairs of loudspeakers show that the confidence interval of the direction estimations tends to be considerably larger in the vertical dimension than in the horizontal one [38, 39, 40, 3, p. 32f].
Pulkki and Wendt et al. also assessed the localization performance when panning with triangles of loudspeakers, used in spatialization methods such as VBAP and AllRAD decoding. In such triple-loudspeaker scenarios, the same effect of different spreads in the median and horizontal plane can be observed. Wendt et al. explain the high zenithal mapping errors with large *intrasubjective* differences, whereas the *intersubjective* standard deviation of repeated trials of the same person is much smaller and comparable to the azimuthal standard deviation [38, 40]. Apparently, the localization performance in the median plane can vary significantly and depends strongly on the individual capabilities of the listeners.
In the further evaluation, azimuthal localization errors of 5° and up to 10° for angular errors in the zenithal dimension are defined as thresholds for an acceptable and plausible localization performance (see also [3, p. 37]). The upper limit of tolerance for the total error e_total, which combines the mismatches in azimuth and zenith, is determined by 10 °, as well. Thresholds in this range are consistent with the human performance in several studies concerning the localization performance of virtual [41] and discrete sound sources [41, 42, 43, 44].

2.2. Direction Estimation of Loudspeakers

Using coincident microphone arrays, the localization estimation of a single sound source is solely based on level differences detected by the single capsules, since coincident microphone arrays try by design to minimize time delays between their membranes.
For the direction of arrival (DOA) estimation of a sound wave, an implementation of the pseudo-intensity vector (PIV), very similar to the one presented by Tervo in the context of the Spatial Decomposition Method (SDM)[24], is applied: The 1st-order Ambisonic impulse response is bandpass-filtered from 200 Hz to 4 kHz in order to avoid poor results caused by interference effects due to diffraction of frequencies with large wavelengths (in relation to the microphone dimension) and spatial aliasing. With this filtered signal, the PIV $\boldsymbol{I}(t)$ is computed:

$$\boldsymbol{I}(t) = W_{\text{filt}}(t) \begin{bmatrix} X_{\text{filt}}(t) \\ Y_{\text{filt}}(t) \\ Z_{\text{filt}}(t) \end{bmatrix} . \tag{2.1}$$

As recommended by Tervo, a moving median window is used to smooth the intensity vector $\boldsymbol{I}(t)$. Since an arriving sound wave reaches the individual capsules with a small time delay, a smoothing window ensures that pressure values, belonging to the same wave front, are collectively assessed.
The length of the window should equal the time period a sound wave needs to propagate through the entire microphone array; a too long window length increases the probability of evaluating multiple incoming wave fronts in a single instance [24, p. 20]. Eq. 2.2 yields the minimum window length in samples dependent from the distance $d_{\max}$ of the two most distant pressure sensors.

$$L_W \geq \frac{2d_{\max}}{c\Delta t} \quad \text{with} \quad \Delta t = \frac{1}{f_s} \tag{2.2}$$

with the sampling rate $f_s = 44.1\,\text{kHz}$, the speed of sound $c = 343\,\frac{\text{m}}{\text{s}}$ and the estimated $d_{\max} = 2\,\text{cm}$ of the Soundfield ST450 MKII, eq. 2.2 yields the window length $L_W \approx 5$ samples.
In the next step, the time instance $t_{\max}$ is detected where the direct sound wave reaches a capsule of the microphone array and therefore the smoothed PIV exhibits its maximal length. For separating the direct part of the impulse response a time interval of 1 ms around the maximum length of the pseudo-intensity vector $\boldsymbol{I}(t = t_{\max})$ is selected and used for estimating the DOA of the incoming direct sound. The selection window is placed asymmetrically around the peak: It starts 0.25 ms before and ends 0.75 ms after $t_{\max}$.
The pseudo-intensity values are summed up and yield a Cartesian vector $\boldsymbol{\theta}_{\text{p},l}$ that points into the estimated direction of the loudspeaker l:

$$\boldsymbol{\theta}_{\text{p},l} = \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \int_{t=t_{\max}-0.25\text{ms}}^{t_{\max}+0.75\text{ms}} \boldsymbol{I}(t) = \int_{t=t_{\max}-0.25\text{ms}}^{t_{\max}+0.75\text{ms}} W_{\text{filt}}(t) \begin{bmatrix} X_{\text{filt}}(t) \\ Y_{\text{filt}}(t) \\ Z_{\text{filt}}(t) \end{bmatrix} dt\,. \tag{2.3}$$

The estimated direction can now be transformed into an expression of azimuth and zenith angle φ and ϑ. The length of the vector can be used as measure for the comprised sound intensity in the respective time interval, but is of no further benefit in the context of this localization method.

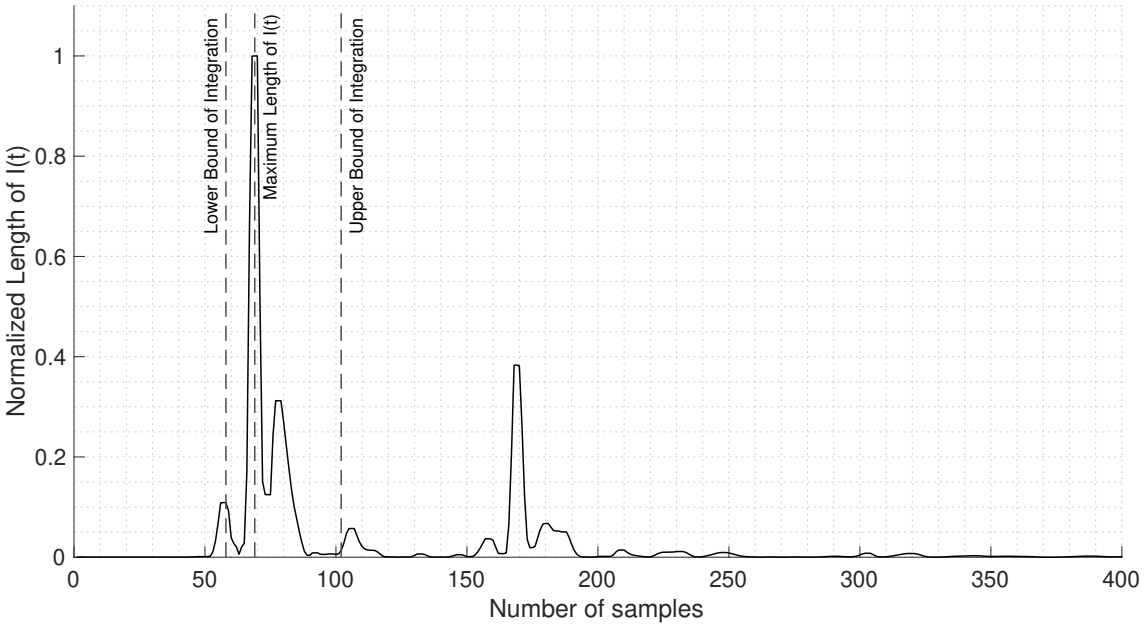


Figure 2.1.: Length of normalized pseudo-intensity vector $\boldsymbol{I}(t)$ of impulse response of LSP 1 to Mic. Pos. 1 in the IEM CUBE and lower and upper bound of the 1 ms long integration interval.

2.3. Direction Estimation of Phantom Sources

The precise prediction of the localization of phantom sources via the energy vector model is attested in a couple of publications [45, 46, 40]. Suitable extensions, which take different parameters such as arrival time, level, direction, and the precedence effect into account, made the model applicable as well at off-center listening positions [47, 48, 49]. In the following paragraph, different vector models are briefly presented and the derivation by Kurz of the vector model, used in this thesis, is explained.

Fundamentals The general form of a vector model, which predicts the perceived source direction at an arbitrary listener position in 3D for a loudspeaker arrangement with L loudspeakers, can be formulated as

$$\boldsymbol{r}_\gamma = \frac{\sum_{l=1}^{L} g_l^\gamma \boldsymbol{\theta}_l}{\sum_{l=1}^{L} g_l^\gamma} \tag{2.4}$$

with the loudspeaker direction vector $\boldsymbol{\theta}_l$ pointing from the investigated listener position to the individual loudspeakers with the particular gains g_l. Depending on the choice of the exponent γ, different slopes of the estimation curve can be achieved: $\gamma = 1$ matches the velocity vector $\boldsymbol{r}_\mathrm{V}$, $\gamma = 2$ results in the energy vector $\boldsymbol{r}_\mathrm{E}$ [3, p. 33] and a direction dependent $\gamma(\varphi)$ was discussed as well in [46].
As described previously, research has shown that the $\boldsymbol{r}_\mathrm{E}$-vector, resp. $\gamma = 2$, proves to be the best fitting predictor in the context of broadband, multi-loudspeaker playback scenarios. Kurz implemented two correction weights w_τ and w_r for the loudspeaker gains and could show that his model works sufficiently accurate for center and off-center listening positions in the context of Ambisonic playback setups [49].

The weight w_r models the sound damping caused by the dissipation of the air, the source is assumed to be point-like. w_r is individually calculated for every loudspeaker l with the normalized distance $\tilde{r}_l = \frac{r_l}{\min(r_l)}$ to the listening position:

$$w_{\mathrm{r},l} = \frac{1}{\tilde{r}_l} = \frac{\min(r_l)}{r_l} \quad . \tag{2.5}$$

The second weight w_τ converts the individual time delays caused by the varying distances to each loudspeaker into level differences:

$$w_{\tau,l} = 10^{\frac{-1000}{4.20}\tau_l} \tag{2.6}$$

with the exponent $\tau_l = \frac{\tilde{r}_l}{c}$. The speed of sound is assumed as $c = 343\frac{\mathrm{m}}{\mathrm{s}}$.
Complementing eq. 2.4 with the presented weights and $\gamma = 2$, we get the definition of the extended model of the energy vector

$$\boldsymbol{r}_{\mathrm{E}} = \frac{\sum_{l=1}^{L}(w_{\mathrm{r},l}\ w_{\tau,l}\ g_l)^2\boldsymbol{\theta}_l}{\sum_{l=1}^{L}(w_{\mathrm{r},l}\ w_{\tau,l}\ g_l)^2} \quad . \tag{2.7}$$

Adaption So far, the $\boldsymbol{r}_{\mathrm{E}}$-model was applied only to known loudspeaker positions in literature [40, 45, 46, 47, 48, 49]. Now, the knowledge of the individual loudspeaker gains g_l, the direction estimation (see sec. 2.2), and the transfer functions themselves provide enough information for the implementation of the extended energy vector model. The loudspeaker gains could also be extracted from the measurement, which is unnecessary in this work, since the gains are known from the signal processing stage in the virtual work environment.
The distance r_l, which is necessary for the calculation of the weights w_r and w_τ, can be extracted from the interval of time, which passes until the first sound wave traveled from the loudspeaker membrane to the microphone position. In order to determine the duration, the maximum length $\max\{\|\boldsymbol{I}(t)\|\}$ of the pseudo-intensity vector is detected. For example, in case of the IEM CUBE and the transfer function from loudspeaker 1 to microphone position 1 (see fig. 2.1), it takes $\Delta_l = 69$ samples (≈ 1.6 ms at a sampling rate $f_s = 44.1$ kHz) until the first wave front arrives. r_l can be calculated with

$$r_l = \frac{c \cdot \Delta_l}{f_s} \tag{2.8}$$

given the speed of sound $c = 343\ \frac{\mathrm{m}}{\mathrm{s}}$ and $f_s = 44.1$ kHz.

2.4. Approaches to a Robust Loudspeaker Localization

A reliable localization of loudspeakers with a microphone array can save a lot of time and effort, since it makes the complicated distance measuring by hand unnecessary. If the measurements should serve as foundation for phantom source localization using the extended $\boldsymbol{r}_{\mathrm{E}}$-vector model, they should be as exact as possible in order to avoid error propagation. The following section proposes some methods, which try to minimize the effort and still yield reliable and accurate localization results.

The simplest measurement approach would be using only a single microphone position for the direction estimation of the loudspeakers. The most suitable location for the microphone array is the center of the room, since it is farthest from the walls and thus deteriorating early reflections can be avoided.

This method is used applying the $\boldsymbol{r}_{\mathrm{E}}$-vector model for phantom source prediction presented above (see sec. 2.3). As the evaluation in sec. 4.2.1.1 shows, this approach yields acceptable results of direction mismatches $\leq 10^{\circ}$ at the central position. With the knowledge of a single loudspeaker position, the evaluation proved that the detection offset could be even further decreased by aligning the estimated direction with the known direction.

Until now, only the estimation of *directions* was discussed. With direction estimations from multiple microphone positions in the vicinity of the central position, it is possible to determine the *position* of a loudspeaker. For example, microphone array positions, arranged in a 1-by-1 m grid around the center, can be combined in a least squares approach. The advantage of this procedure is that singular direction mismatches can be compensated in combination with other measurements. Preliminary tests and evaluations of the author showed promising results but could not be further pursued due to the limits of this thesis. Questions concerning the optimal ratio of number of measurements and achievable accuracy are yet to be answered. The necessary, theoretical foundations for this methodological approach are presented in the appendix (see app. A.2).

3. Quality Measures for Surround Sound Setups

This chapter is concerned with the development and adaption of quality measures for the assessment of multi-loudspeaker systems. A crucial feature for the quality of a loudspeaker setup is its ability to synthesize virtual sound sources appropriately. The focus of this chapter lies on the robustness and the easy application of the developed measures, rather than designing models that derive measures by describing the psycho-acoustic mechanisms of the ear.
Static, spatially invariant phantom sources need to be localized accurately and the direct sound should always be more present at the position of the listener than the diffuse sound reflected by room surfaces. Moving phantom sources add further and higher demands on loudspeaker systems regarding the stability of features, such as sound level, spectral envelope, and apparent source width (ASW).
The first section 3.1 presents the development environment, designed for the virtual evaluation of the quality measures. In sec. 3.2, the direct-to-reverberation energy ratio (DRR) is utilized to confine the sweet area of a loudspeaker setup. The lateral energy fraction (LF) showed to be a valid indicator for the spread of phantom sources and is optimized in section 3.3 for the application at multi-loudspeaker systems. Section 3.4 concentrates on the measure C_m for spectral stability (coloration) of phantom sources. The focus of the last section 3.5 lies on the novel measure EV_m for the envelopment of the listener (LEV), achieved by the sound reproduction system.
The adapted and newly developed measures can be recognized by their index m, since they are all adapted for the application at $m$ulti-loudspeaker audio reproduction setups.

3.1. Virtual Working Environment

The transfer function from a loudspeaker to a microphone in a room is determined by measuring the impulse response of the room using the loudspeaker as playback and the microphone as recording device. Thereby, the performance of a surround sound system reproducing spatial audio can be fully described as a linear system by measuring the transfer functions from each loudspeaker to one or more microphones positioned in the listening area. Through the convolution of a loudspeaker signal with its corresponding transfer function, it is possible to compute exactly how the loudspeaker signal would sound recorded at the microphone position. Since it is a linear system, also the resulting microphone signal of the superposition of multiple loudspeaker signals can be computed. Any spatialization method can be virtually simulated on the foundation of the gathered transfer functions with one full measurement procedure including all loudspeaker and microphone positions.
All used measurements in this thesis were performed with a compact microphone array, which can capture 1st-order Ambisonic room impulse responses. This allows to incorporate further measures into the evaluation, which are based on the directional analysis of sound signals.
The virtual working environment for the quality measures in this thesis can be divided into three stages (see fig. 3.1): In the first stage, the *input stage*, the signal, the impulse

response measurements, as well as the spatialization method and its related parameters are defined. In the subsequent *signal processing stage*, the loudspeaker signals are rendered and convolved with their corresponding transfer functions. The resulting 1st-order Ambisonic microphone signal for the final *output stage* is computed by summing the individual, convolved signals and serves as foundation for further investigations.
The quality measures in this chapter were developed on basis of 720 1st-order Ambisonic impulse responses from 24 loudspeakers to 30 microphone positions measured with two microphone arrays of the model Soundfield ST450 MKII in the IEM CUBE (see sec. 4.2) [50].

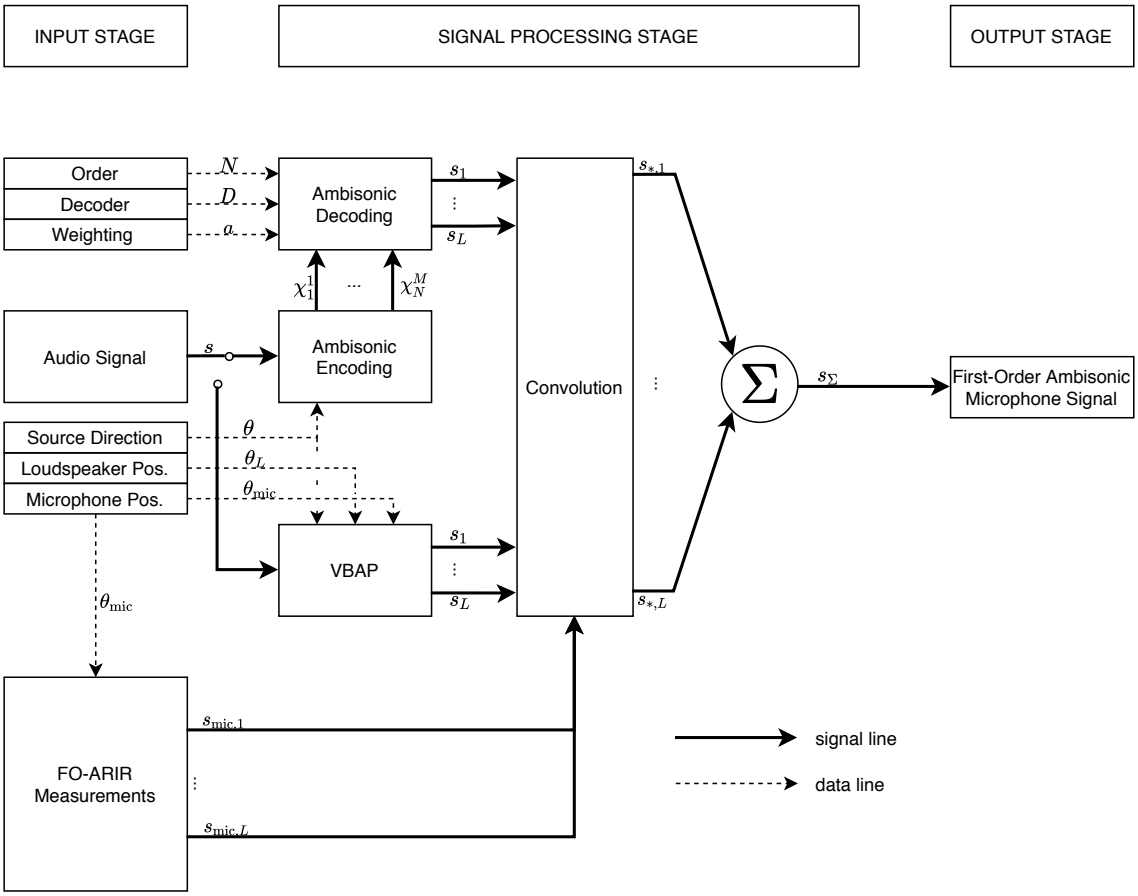


Figure 3.1.: Signal flow diagram of the virtual working environment utilized in this work.

3.2. Adapted Direct-to-Reverberant Energy Ratio (DRR_m)

Fundamentals The conventional direct-to-reverberant energy ratio DRR is a logarithmic measure for monophonic, omni-directional impulse responses. It is determined by the ratio of the direct sound energy, which is comprised in the first part of the impulse response and the sound energy of the second, reverberant part of the impulse response, which consists of early and late reflections of the room.
The DRR is commonly used as indicator for distance perception of auditory events [42, 51, 52] and describes the influence of the reverberation on the quality of spatial perception of the listeners [53, 54]. Additionally, it can be interpreted with focus on the direct sound energy, which should be predominant at all listening positions in the sweet area ($DRR \geq 0\,\mathrm{dB}$).
For impulse response measurements $s(t)$ of a single discrete sound source exciting the room, it is rather simple to define intervals of direct and diffuse sound. Usually, the direct sound part is detected via the global pressure maximum and windowed from $t_{\mathrm{w},1}$ to $t_{\mathrm{w},2}$ ($t_{\mathrm{w},1} < t_{\mathrm{w},2}$). A wide range of different window sizes from 1.5 ms to 16 ms can be found in literature. Also the exact alignment of the window varies: Often it is applied directly after the maximum peak of the impulse response, but there are also implementations which place the point of onset adequately early before the maximum in order to prevent loss of direct sound energy [52].
Unaffected by the individual window definitions, the general equation for the DRR is defined by

$$DRR = \frac{E_{\mathrm{dir}}}{E_{\mathrm{rev}}} = 10 \log_{10} \left(\frac{\int_{t_{\mathrm{w},1}}^{t_{\mathrm{w},2}} s^2(t) dt}{\int_{t_{\mathrm{w},2}}^{\infty} s^2(t) dt} \right) . \tag{3.1}$$

Adaption Phantom sources on multi-loudspeaker setups are synthesized with VBAP or Ambisonics by several active loudspeakers. Depending on the position of the microphone, direct sound parts from the individual loudspeakers arrive with different time delays. The direct and diffuse parts of the loudspeaker signals partially superpose each other and are therefore perceived collectively.
In the following, the direct-to-reverberant energy ratio is adapted for the assessment of multi-loudspeaker systems: Now, the omnidirectional channel $W(t)$ of the computed 1st-order Ambisonic microphone signal corresponds to $s(t)$. It consists of the sum of the measured impulse responses, multiplied with the gains of the loudspeaker that synthesize a phantom source. We define the direct sound energy as the energetic content, which is included in the time interval from the beginning ($t_0 = 0\,\mathrm{ms}$) of the impulses response until t_1, the end of the lag d_{lag} after the loudest peak at $t_{\max}$ of the squared sound pressure $W^2(t)$. Afterwards, a cross-fade section of duration d_{cross} leads over to the diffuse part of the impulse response. Mathematically, the cross-fade is implemented through the multiplication with the two weighting functions $w_{\mathrm{dir}}(t)$ for the direct and $w_{\mathrm{rev}}(t)$ for the diffuse part.

They are specified as

$$w_{\mathrm{dir}}(t) = \begin{cases} 1 & \text{if } 0 \leq t \leq t_1 \\ 1 - \frac{1}{d_{\mathrm{cross}}}(t - t_1) & \text{if } t_1 < t \leq t_2 \\ 0 & \text{if } t > t_2 \end{cases} \tag{3.2}$$

and

$$w_{\mathrm{rev}}(t) = \begin{cases} 0 & \text{if } 0 \leq t \leq t_1 \\ \frac{1}{d_{\mathrm{cross}}}(t - t_1) & \text{if } t_1 < t \leq t_2 \\ 1 & \text{if } t > t_2 \end{cases} \tag{3.3}$$

with $t_1 = t_{\max} + d_{\mathrm{lag}}$ and $t_2 = t_{\max} + d_{\mathrm{lag}} + d_{\mathrm{cross}}$.
The adapted direct-to-reverberant energy ratio can now be defined as

$$DRR_m = 10 \log_{10} \left(\frac{\int_0^{t_2} W^2(t) w_{\mathrm{dir}}(t) dt}{\int_{t_1}^{\infty} W^2(t) w_{\mathrm{rev}}(t) dt} \right) . \tag{3.4}$$

In order to make valid statements about the accuracy and accordance to the psychoacoustic perception of this measure, further listening tests have to be performed (see sec. 3.6).
In the meanwhile, some educated guesses for d_{lag} and d_{cross} can be made: From informal listening experiences, the author knows that static phantom sources decoded in 7th-order Ambisonics equal strongly the impression of a sound source represented by a discrete loudspeaker. The two values were chosen in a way, so that the graphic evaluation of these two variants are similar to each other.

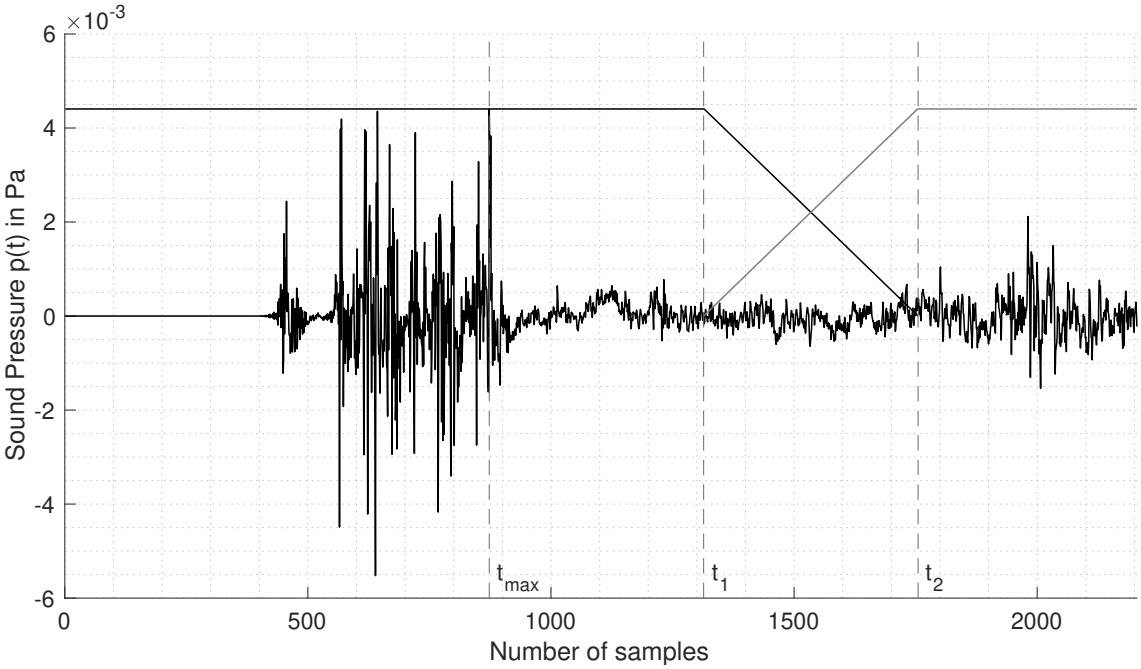
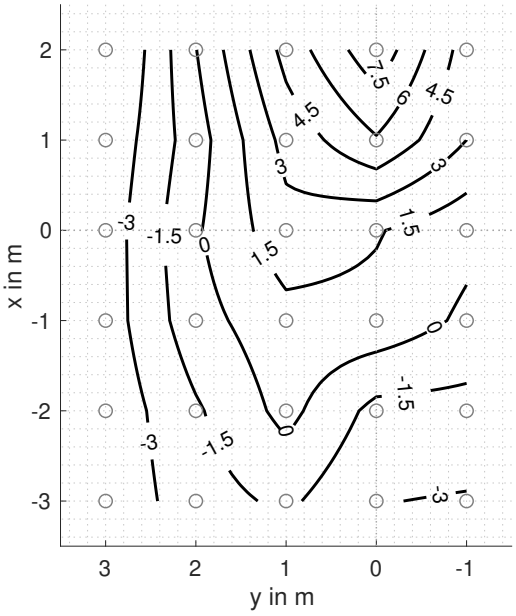


Figure 3.2.: W-channel the microphone signal and envelopes of the weighting functions $w_{\mathrm{dir}}(t)$ (black) and $w_{\mathrm{rev}}(t)$ (grey).

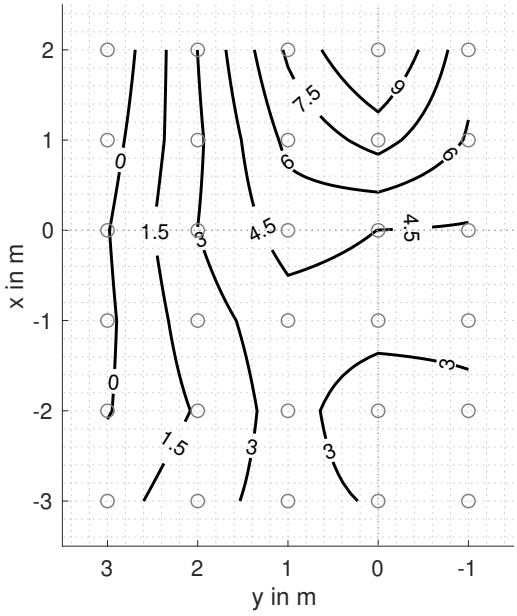
Fig. 3.2 shows the W-channel of a simulated microphone signal of a phantom source, which is decoded in 3rd-order Ambisonic at the position of loudspeaker 1 in the IEM CUBE. d_{lag} and d_{cross} equal 10 ms in this figure. The weighting functions w_{dir} and w_{dif} are marked as black and grey lines in the graph. Their exact values do not correspond to the figure,

as they only should clarify the temporal separation and the cross-fade from the direct to reverberant part of the impulse response. Although the phantom source is exactly at the position of a loudspeaker, the 3rd-order Ambisonics is just able to yield a rather wide source directivity pattern. This leads to a participation of multiple loudspeakers, which is recognizable by the numerous peaks until time instance $t_{\max}$.

Comparison The design of the standard DRR does not consider superimposed direct parts from multiple loudspeakers; delayed direct sound energy from distant loudspeakers is not assessed, since it arrives outside of the fixed evaluation window.
Since the limits proved to be very suitable, the same window as in the direction estimation implementation is chosen for the DRR window (see sec. 2.2): The direct sound interval starts 0.25 ms before the peak of the length of the PIV and lasts until 0.75 ms thereafter. Findings by Csadi et al. support this choice and suggest that window lengths below 2.5 ms in rather small rooms yield more consistent results [52]. In order to assess exhibited discrepancies, the standard DRR and DRR_m are compared in two scenarios.
As observable in Fig. 3.3, both measures exhibit similar contours for the single active loudspeaker 1. The lower values of the conventional DRR are caused by the shorter time window and the therefore lower energetic level.
The limitations of the standard measure become visible by analyzing phantom sources, synthesized by multiple loudspeakers. Fig. 3.4 presents the two energy ratios for a 7th-order Ambisonic phantom source decoded in the middle between loudspeaker 1 and 2 ($\boldsymbol{\theta} = [\text{-}11.02°, 90°]$). While the classic measure DRR collapses disproportionally strong in the left boundary area, the adapted DRR_m, apart from a slight shift to the right, exhibits comparable pattern to fig. 3.3b. Hence, the adapted measure proves to describe the subjective perception more suitably. The intense decrease on the left boundary of the standard measure is caused by the growing time delays of the active loudspeakers at the more distant microphones positioned towards the left side: As shown in fig. 3.5a, the direct part from loudspeaker 2 arriving at microphone position 20 is outside of the standard DRR evaluation window. The adapted definition of the window in the implementation of the DRR_m resolves this malfunction (see fig. 3.5b).

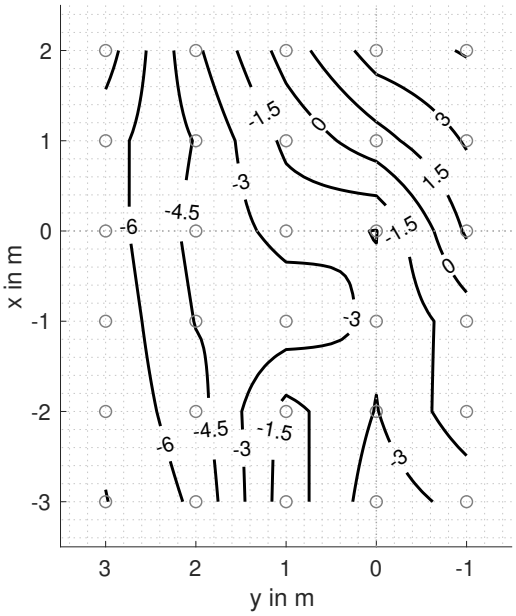


(a) Standard DRR measure.

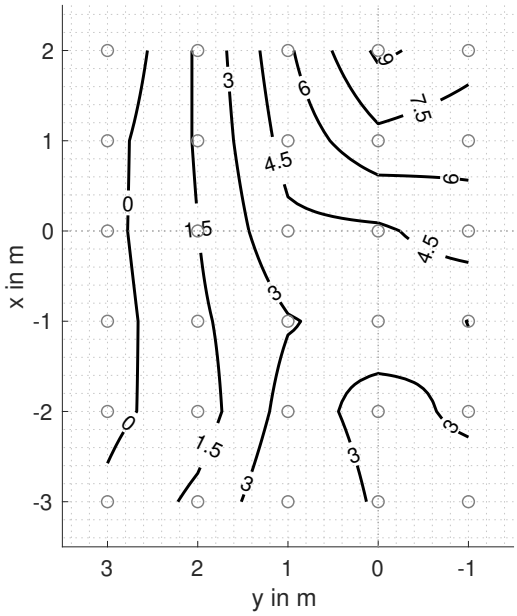


(b) Adapted DRR_m measure.

Figure 3.3.: Standard and adapted evaluation in dB of a single discrete sound source at LSP 1 ($\boldsymbol{\theta}_\mathrm{p} = [4.34, 0.12, 1.34]$).

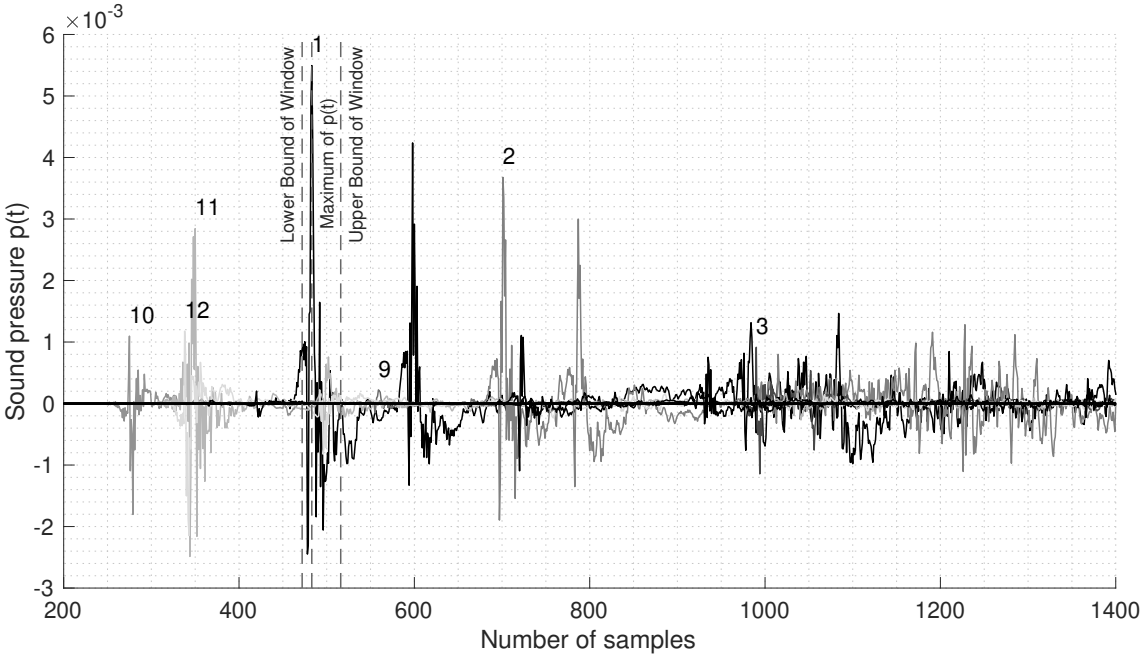


(a) Standard DRR measure.

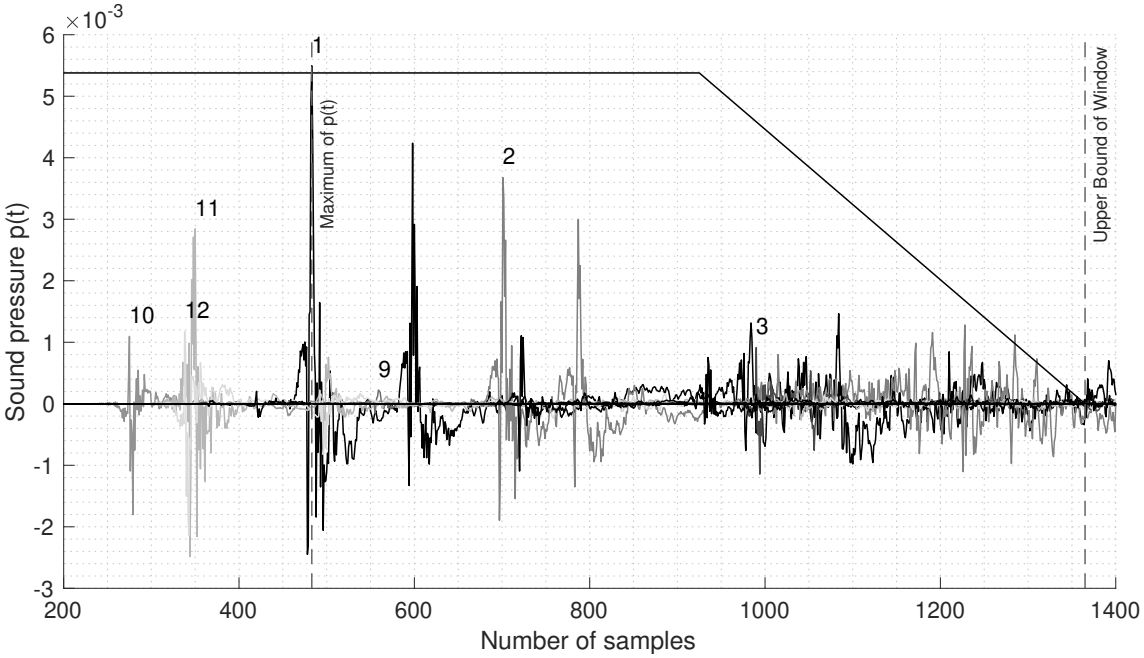


(b) Adapted DRR_m measure.

Figure 3.4.: Standard and adapted evaluation in dB of a phantom source encoded in 7th-order Ambisonics between LSP 1 and 2.



(a) Standard DRR measure.



(b) Adapted DRR_m measure.

Figure 3.5.: Integration windows and individual loudspeaker signals marked with their numbers computed for mic. pos. 20.

3.3. Adapted Lateral Energy Fraction (LF_m)

Fundamentals Several publications show that the conventional lateral energy fraction LF is a good indicator for the perceived spread of an auditory event, also known as the *apparent* or *auditory source width* (ASW) [45, 55, 56]. Hidaka et al. calculate the LF through the ratio of the energy of the lateral reflections in the time interval from 5 to 80 ms to the energy of the entire signal from 0 to 80 ms. The reference signal $s_\circ(t)$ is recorded with an omnidirectional microphone, whereas the lateral reflections $s_\infty(t)$ are measured with a microphone, which exhibits a figure-of-eight characteristic [56]:

$$LF = \frac{\int_{5\text{ms}}^{80\text{ms}} s_\infty^2(t)dt}{\int_{0\text{ms}}^{80\text{ms}} s_\circ^2(t)dt} \,. \tag{3.5}$$

The LF measurement can be performed with a mid-side (M/S) recording setup as shown in fig. 3.6.

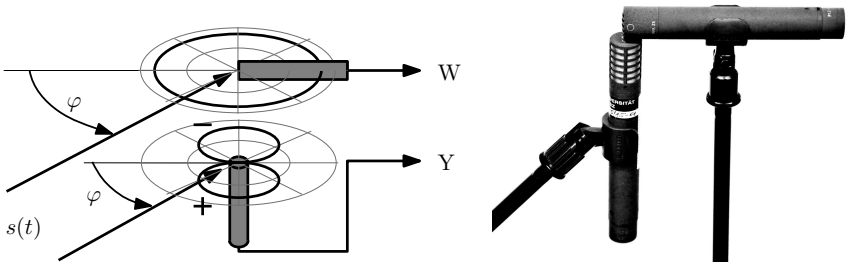


Figure 3.6.: Schematic mid-side recording setup on the left side and the corresponding arrangement of two microphones on the right, from [3, p. 5].

Adaption The lateral energy fraction LF can be adapted for the evaluation of 1st-order Ambisonic room impulse responses. As fig. 3.6 implies, the omnidirectional signal can be interpreted as the W-channel and the pattern-of-eight signal as the Y-channel in a B-format recording of a first-order microphone array. The use of FO-ARIRs enables the possibility to change virtually the orientation of the M/S-setup after the measurement is done. This feature proves to be highly useful, especially if orientation offsets, which are noticed during the evaluation, have to be equalized or multiple LF estimations in different directions shall be assessed with a single recording. The orientation of the Y-channel can be rotated around the z-axis to every desired azimuthal angle $\tilde{\varphi}$ with a suitable rotation matrix. Since the spherical harmonics Y_n^m of degree $m = 0$ are axially symmetric in relation to the vertical z-axis, only the two spherical harmonics Y_1^{-1} and Y_1^1, corresponding to the Y- and X-channel, have to be manipulated. The azimuthal offset angle φ_Δ to the desired direction $\boldsymbol{\theta}$ [3, p. 114ff] is calculated with

$$\varphi_\Delta = \tilde{\varphi} - \varphi \,. \tag{3.6}$$

The manipulation is expressed by

$$\begin{bmatrix} \tilde{Y} \\ \tilde{X} \end{bmatrix} = \begin{bmatrix} Y_1^{-1}(\varphi + \varphi_\Delta, \vartheta) \\ Y_1^1(\varphi + \varphi_\Delta, \vartheta) \end{bmatrix} = \boldsymbol{R}(m, \varphi_\Delta) \begin{bmatrix} Y_1^{-1}(\varphi, \vartheta) \\ Y_1^1(\varphi, \vartheta) \end{bmatrix} = \boldsymbol{R}(m\varphi_\Delta) \begin{bmatrix} Y \\ X \end{bmatrix} \tag{3.7}$$

with

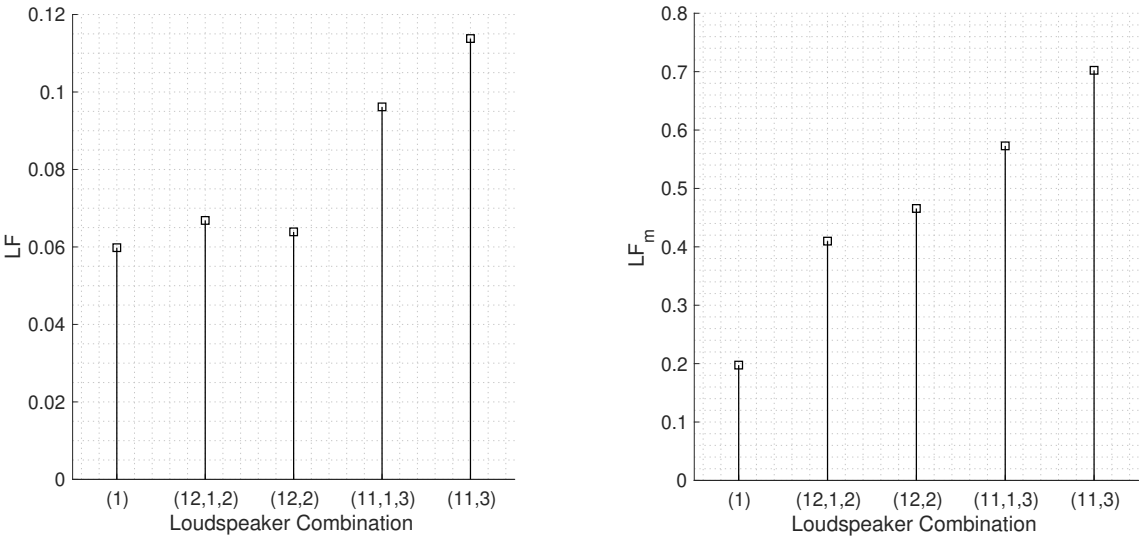
$$\boldsymbol{R}(m, \varphi_\Delta) = \begin{bmatrix} \cos(m\varphi_\Delta) & \sin(m\varphi_\Delta) \\ -\sin(m\varphi_\Delta) & \cos(m\varphi_\Delta) \end{bmatrix}, \tag{3.8}$$

and the Ambisonic degree $m = |\pm 1|$. This operation also can be understood as rotating the spherical coordinate system relative to the spherical harmonics around the z-axis until the x-axis points in the same direction as $\boldsymbol{\theta}$. The zenithal angle ϑ of $\boldsymbol{\theta}$ can be omitted, since it would yield a rotation around the y-axis, to which the directivity patterns of the W-channel (Y_0^0) and the $\tilde{Y}$-channel ($\tilde{Y}_1^{-1}$) are rotation-symmetric.
Furthermore, Frank proposes that the prediction of the ASW with the LF improves by decreasing the lower integration bound of the figure-of-eight signal from 5 ms to 0 ms in the context with the application to multi-loudspeaker setups. His regression analysis shows that the conventional LF relates to results of listening tests very poorly ($R^2 = 0.19$), whereas the adaption performs significantly better ($R^2 = 0.69$) [57].
Combining all the presented details, the adaption of the lateral energy fraction for the virtual working environment is given by

$$LF_m(\tilde{\varphi}) = \frac{\int_{0\text{ms}}^{80\text{ms}} \tilde{Y}^2(t)\, dt}{\int_0^{80\text{ms}} W^2(t)\, dt} = \frac{\int_{0\text{ms}}^{80\text{ms}} (Y(t)\cos(\varphi_\Delta) + X(t)\sin(\varphi_\Delta))^2\, dt}{\int_0^{80\text{ms}} W^2(t)\, dt}. \tag{3.9}$$

The interval of possible LF_m values is defined between [0,1]: A very wide source yields a high level of lateral sound energy contained in the $\tilde{Y}$-channel; however, the energy level can never exceed the energy content of the omnidirectional channel $W(t)$. If a very narrow source is positioned in the zero of the figure-of-eight pattern of $\tilde{Y}$, it follows that $\int_{0\text{ms}}^{80\text{ms}} \tilde{Y}^2(t)\, dt = 0$. The value of LF_m is proportional to the widening of the ASW from a very narrow source ($LF_m = 0$) to its maximal spread ($LF_m = 1$).

Comparison In this paragraph, the differences between the standard LF and the adapted LF_m measure are examined. Evaluating from microphone position 17 the ASW of a discrete source positioned at loudspeaker 1, the standard measure $LF = 0.06$ (see fig. 3.7a) is considerably smaller than the revised measure $LF_m = 0.20$ (see fig. 3.7b). The energy of the direct part comprised the first 5 ms of the impulse response is not included in the integral in the nominator and therefore the entire fraction is smaller.
Using the measurements conducted in the IEM CUBE, both measure variants were evaluated for the following combinations of one, two, and three frontal loudspeakers: (1), (12,2), (11,3), and (12,1,2), (11,1,3) (see tab. 3.1). Loudspeaker 1 is positioned centrally in front of the microphone. The microphone position and the positions of the remaining loudspeakers can be reviewed in fig. 4.2. According to the results of Frank's listening experiment in [57], the ASW should rise with the increase of the aperture angle of the loudspeaker basis and the addition of a center loudspeaker should decrease the perceived source spread. The conventional LF measure exhibits inconsistent and contradictory values for the different combinations (see fig. 3.7a). For example, the addition of the center loudspeaker to the pair 1 and 2 increases the measure and only a small perceptual difference to a discrete sound source at LSP 1 is indicated.
In contrast to that, fig. 3.7b shows that the adapted measure meets the expected course of the lateral energy fraction values. In tab. 3.1, the aperture angles of the three investigated pairs of loudspeakers can be found.



(a) Standard LF measure.

(b) Adapted LF_m measure.

Figure 3.7.: Standard and adapted evaluation of different LSP combinations on the surround sound setup in the IEM CUBE.

Pair of LSPs	(12,2)	(11,3)
Aperture angle α in $^\circ$	47.9	97.1

Table 3.1.: Aperture angle of the three investigated pairs of loudspeakers.

Error Estimation When determining the ASW of a discrete source without the knowledge of the measured position of the loudspeaker, the orientation of the LF_m is dependent on the DOA estimation presented in sec. 2.2.
An error estimation was conducted to investigate how strongly an azimuthal offset φ_Δ impacts the LF_m evaluation of loudspeaker 1 at the central microphone position. Since it is recommended to use the central microphone position for the direction estimation, the influence of the mapping errors at this position is of interest. Loudspeaker 1 shows the biggest error at microphone position 17 compared to the other horizontal loudspeakers for $\varphi = [0°, 17°]$. The error is defined by the ratio of the biased $LF_m(\varphi_1 + \varphi_\Delta)$ to the LF_m at the measured position of loudspeaker 1:

$$e_{LF_m}(\varphi_\Delta) = 10 \log_{10} \left(\frac{LF_m(\varphi_1 + \varphi_\Delta)}{LF_m(\varphi_1)} \right) . \tag{3.10}$$

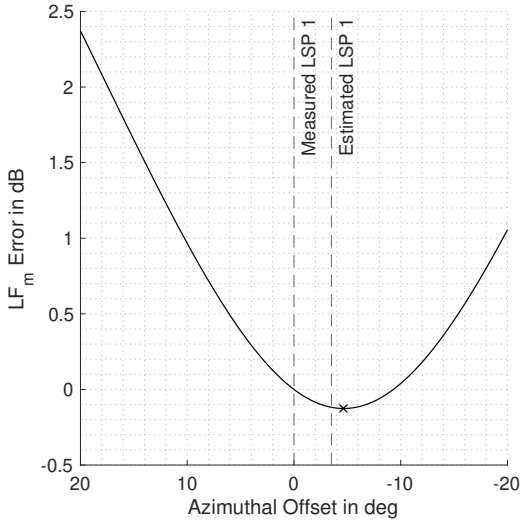


Figure 3.8.: Error $e_{LF_m}(\varphi_\Delta)$ in dB for LSP 1 at mic. pos. 17.

Although the minimum is usually expected at the measured loudspeaker position, the minimum of e_{LF_m} (marked with $\times$ in fig. 3.8) lies on the right side of the measured and estimated LSP position. This could be explained by localization inaccuracies, both measured or estimated or the inaccurate figure-of-eight pattern of the microphone. Also, the spatial conditions of the mounting position of LSP 1 in the IEM CUBE can provoke reflections, which deteriorate the ideal result.
As shown in fig. 3.8, the detected median azimuthal offset of loudspeakers 1 to 12 of $e_{\text{total}} \approx 5\,°$ (see sec. 4.2.3, tab. 4.1) causes deviations of less than 0.5 dB, which would be equivalent to an offset of approx. 0.019 to the correct LF value $LF_m \approx 0.2$ of loudspeaker 1 in absolute numbers. An offset of 20 ° translates to $LF_m \approx 0.35$ for loudspeaker 1 in the surround sound setup of the IEM CUBE.
The JND of the LF equals roughly 0.05 [58, 59]. Further evaluation results, presented in app. A.3, classify the presented LF_m errors in connection to other combinations of microphone positions and loudspeakers.

3.4. Measure for Differences in Coloration (C_m)

Reproducing phantom sources with multiple loudspeakers generates unavoidable comb-filters at the listener's position, induced by the superposition of coherent signals. Especially in the context of moving sound sources, prominent spectral fluctuations can be very problematic, since they are time-variant, dependent on the individual listener's position and thus are unequalizable. Rumsey et al. showed that the timbral fidelity has a dominant influence on the perceived overall quality of a playback system [60]. Therefore, a measure for the audibility of spectral fluctuations is a crucial parameter for determining the quality of a loudspeaker setup.

Fundamentals Frank conducted listening experiments and measurements with step-wise moving phantom sources on a horizontal ring of 8 and 16 loudspeakers, respectively [45, p. 83ff]. Test subjects had to rate the changes in coloration of sources synthesized with different spatialization methods, including Ambisonics decoding using max-$\boldsymbol{r}_\mathrm{E}$ weighting and VBAP, in steps of 5 ° of the azimuthal angle φ. The results of the measurement were assessed by calculating the differences Δ_dB between the sound levels of the third-octave band filtered signals $s_i(t)$ and $s_{i+1}(t)$ of two neighboring steps i and $i+1$. Frank's evaluation of the results indicated that the median ratings in the experimental setting correlates with $\rho = 0.90$ at the central listening position and with $\rho = 0.71$ at an off-center listening position to the strongest spectral fluctuation at each step. Based on the results of Karjalainen, differences of less than 1 dB in spectrum are treated as not perceivable in this work [61].

Adaption Frank's evaluation method of coloration is adopted as a quality measure for surround sound setups in this work. The third-octave bands b of the filter bank for the spectral assessment of the signal had a quality $Q = \infty$ and were scaled in accordance with EN 61260-1:2014 [62]. The first 35 ms after the first peak in W-channel of the simulated microphone signal is used as foundation for the calculation of C_m. As an enhancement to Frank's setup, not only the horizontal ring, but all loudspeakers of the hemispherical arrangement are available and participate depending on the signal processing method and phantom source position at the sound reproduction. The single number value for the degree of coloration is defined by

$$C_m = \max\{\Delta_\mathrm{dB}(s, b)\}\,. \tag{3.11}$$

3.5. Measure for Envelopment (EV_m)

Probably the most fascinating innovation of surround sound systems for spatial audio is the possibility of giving the listeners the subjective perception of *being surrounded by sound* [63]. As follows, an authentic *envelopment* of the listener is a vital characteristic of any spatial audio reproduction setup. The research focused on the listener's envelopment in the context of multi-loudspeaker audio systems is rather young and to the knowledge of the author, a simple and robust measure could not be established yet.

Fundamentals In the research field of concert hall acoustics, consolidated concepts of describing spatial impressions are already existing: It is regarded as undisputed that the spatial impression of a sound event is strongly dependent on (i) its apparent/auditory source width, as well as on (ii) its lateral reflections surrounding the listener, excluding the direct sound that is coming from the sound source [64]. The influence of the remaining sound field on the spatial impression is widely named *Listener Envelopment* (LEV) in publications [55, 65, 66, 67]. Originating from this traditional use, the term was already adapted regarding the description of multi-loudspeaker audio systems [68]. In this work, the shortened term *Envelopment* (EV_m) is employed for the newly developed quality measure.
A sharp distinction should be made to the expression *Immersion*, which also describes a kind of surroundedness of sound, but more in the field of virtual reality (VR) and artificial soundscapes. In addition, it stresses more the fact of *interactively being* in an environment, rather than feeling surrounded by something [63].

Adaption This paragraph proposes a new, experimental measure for the perception of envelopment, which is primarily based on the investigation of level imbalances. The underlying assumption for this measure is that in order to feel enveloped by a sound reproduced on a hemispherical loudspeaker system, the (direct) sound has to arrive equally distributed from all around at the listener's position, similar to the impression in an ideal diffuse field. Investigations by Hiyama indicate that at least six loudspeakers placed in even intervals of 60 ° on a circle are needed in order to reproduce an authentic spatial impression of a diffuse field with surround sound setups in a horizontal plane, where no front or back direction is defined [69]. It seems plausible to use this aperture angle as evaluation scope, since the perception of diffuse sound fields and envelopment are tightly interconnected.
For the implemented evaluation, the multi-loudspeaker setup is divided into spherical wedges with an aperture angle $\alpha = 60°$ (see fig. 3.9), each shifted by the step-size of 1°. For every loudspeaker, the sound energy, which arrives in the first 35 ms of the signal at the specific microphone position, is calculated. The specific length of the time window was chosen in accordance with the time weighting option called *Impulse*, which was defined in the superseded IEC60651:1979 standard [70] and is still implemented in almost all sound level meters, including the *NTi XL2*[1]. The loudspeaker signals are computed by convolving the transfer function with a Dirac impulse of height 1. The energy levels of loudspeakers, which lie in the shared spherical wedge, are summed and consecutively logarithmized. In that way, the directional measure $EV_m(\varphi)$ is developed, which yields the coverage of sound energy with the directional resolution of $\Delta_\varphi = 1°$ at the assessed microphone position.

[1] https://www.nti-audio.com/en/products/xl2-sound-level-meter.
An explanation of the used time windows can be found under https://www.nti-audio.com/en/support/know-how/fast-slow-impulse-time-weighting-what-do-they-mean.

The calculation method can be formulated as

$$EV_m(\varphi) = 10 \log_{10} \left(\sum_{s_i \in A} \int_{0\text{ms}}^{35\text{ms}} s_i^2(t) dt \right) \tag{3.12}$$

with the set A containing the generated signals $s_i(t)$ of the loudspeakers positioned inside the area $[\varphi - \frac{\alpha}{2}; \varphi + \frac{\alpha}{2}]$, covered by the spherical wedge oriented in the direction φ. The standard deviation $\sigma(EV_m)$ over all spherical wedges can be analyzed in order to determine the strength of the EV_m fluctuation. In addition, the average $\overline{EV}_m$ over all directions indicates, how dominant the direct sound is at the position of the listener.

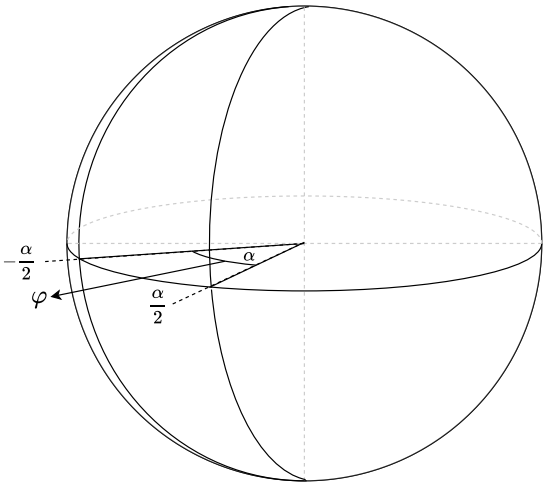


Figure 3.9.: Spherical wedge with aperture angle α.

Fig. 3.10 shows the schematic evaluation with spherical wedges of an exemplary horizontal 8-loudspeaker system with the aperture angle $\alpha = 60°$ and a stepsize 1°.

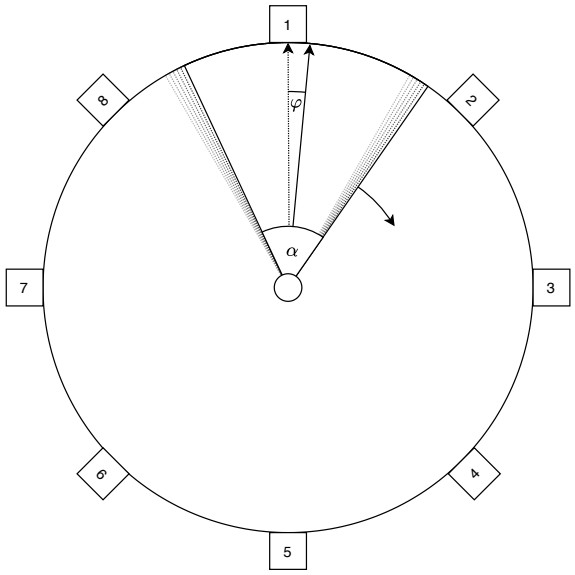


Figure 3.10.: Evaluation of an exemplary playback setup at a microphone position in the center.

3.6. Verification of Measures through Listening Experiments

The detection of phantom sources using the energy vector $\boldsymbol{r}_E$ and the psycho-acoustic measures LF_m and C_m are supported by the results of already conducted listening experiments presented in each section (see sec. 2.3, 3.3 and 3.4).
Since the author could not find supporting results, the two remaining measures DRR_m and EV_m were developed with focus on the adaptability to further findings of listening experiments:
The Direct-Diffuse Ratio DRR_m provides a straightforward and robust framework, which can be adapted by choosing suitable values for the time interval, following d_{lag}, the maximum peak of the impulse response, and the duration d_{cross} of the crossfade between direct and diffuse part.
As well, the Envelopment EV_m could not be verified through listening tests, due to the limits of this work. The aperture angle α of the evaluation wedge, the step-size, and the interval of the impulse response part are parameters that can be optimized in order to fit results of future listening tests. Further, interesting points of research would be the determination of psychoacoustics just noticeable differences (JNDs) for the fluctuation or level notches of the energy levels. At which limits do they start to impair the perceived envelopment? Results of listening experiments conducted at the IEM, which are not yet published, support these perspectives for further research.

4. Evaluation of Different Surround Sound Setups

In the following sections, the two spatialization methods VBAP and AllRAD, presented in chap. 1, are used to produce phantom sources. The AllRAD decoder and the convex hull of VBAP are computed based on the measured loudspeaker positions from [50].

4.1. Preparation of Measurements

All measurements in the three facilities were executed with a Soundfield ST450 MKII microphone, whose preamplifier exports the recordings as signals in first-order Ambisonic B-format.

Microphone Positions In the IEM CUBE and in the György-Ligeti-Hall, a grid of microphone position was measured. Generally, a grid covering the whole listening area is reasonable in order to make valid statements about the boundaries of the sweet area. The denser the grid is laid out, the better the exact limits can be defined. In the context of the direct-to-reverberant energy ratio assessment, a sparse grid can be improved by interpolating linearly between the microphone positions.
In the Production Studio, only one measurement at the central microphone position was performed, since it is optimized for a single audio professional working in the sweet spot of the loudspeaker setup.

Formatting The first step of preparing the measurements for further calculations was to convert the gathered B-format impulse responses $\boldsymbol{s}_{\mathrm{B}}(t)$ into ACN and equalize the weighting of the W-channel by multiplying it with $\sqrt{2}$. This can be done with the simple matrix multiplication [20]

$$\boldsymbol{s}_{\mathrm{ACN}}(t) = \begin{bmatrix} \tilde{W}(t) \\ Y(t) \\ Z(t) \\ X(t) \end{bmatrix} = \begin{bmatrix} \sqrt{2} & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} W(t) \\ X(t) \\ Y(t) \\ Z(t) \end{bmatrix} = \boldsymbol{A}\boldsymbol{s}_{\mathrm{B}}(t) \; . \tag{4.1}$$

Level Equalization The assessment of the IEM CUBE measurements revealed that the impulse responses of the different loudspeakers had varying energy levels in the direct sound part. The author suspects that the amplifiers of the passive loudspeaker in the IEM CUBE caused this bias, since it is a known shortcoming that they adjust to a slightly different gains each time after turning off and on again. However, during a non-interrupted session, the gains stay at a constant offset.
Assuming that all loudspeakers should have the same sound level at the center of the listening area in the IEM CUBE, the measurements were adjusted in a way that the direct part (0.25 ms before and 0.75 ms after the maximum length of the PIV-vector) of all loudspeaker impulse responses to the central microphone position 17 had the same level as the quietest one.

Fig. 4.1 shows the sound energy levels of the direct parts and the first 100 ms of the impulse responses, before and after the energy level correction. All energetic levels were adjusted to the impulse response of loudspeaker 1. The diffuse sound part degrades the strong energetic differences exhibited in the direct part of the impulse response, but still causes a maximal fluctuation between 2.21 dB (LSP 11) and -1.27 dB (LSP 23) in the 100 ms-intervall of the corrected impulse responses. The energetic differences of the uncorrected impulse responses range in an interval from 0 to over 12 dB between loudspeaker 1 and loudspeaker 23.
For the measurements in the György-Ligeti-Hall and the Production Studio no gain compensation was performed. A detailed level analysis for the two loudspeaker setups can be found in the appendix A.4.1.

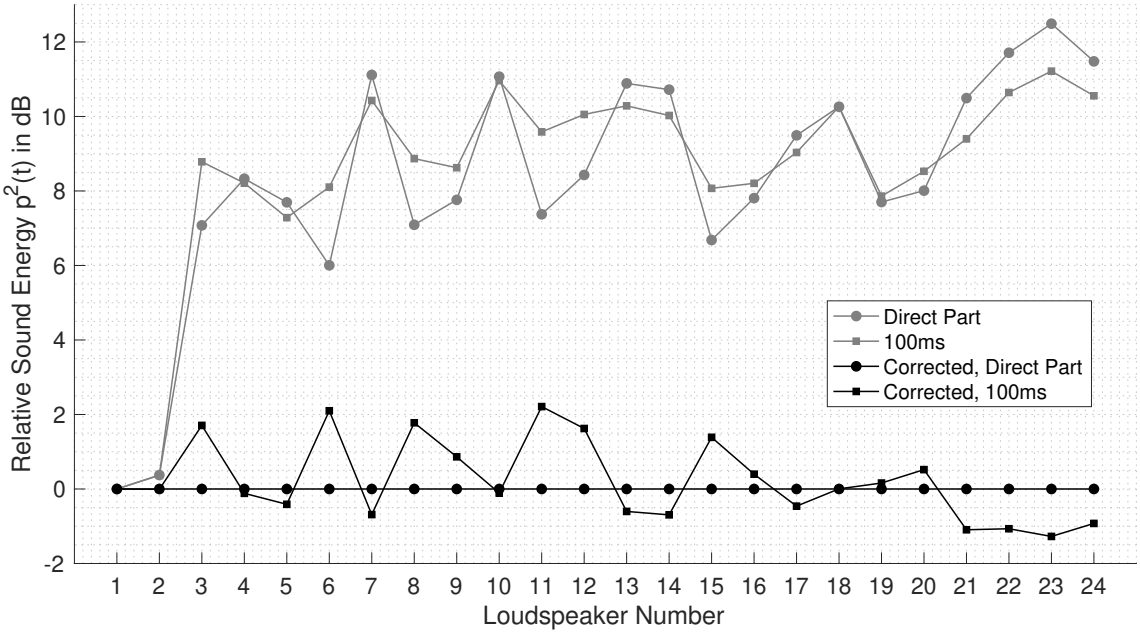


Figure 4.1.: Sound energy of the direct part and the first 100 ms of the uncorrected and corrected loudspeaker impulse responses in the IEM CUBE.

4.2. Full Evaluation of a Lecture Room: IEM CUBE

The IEM CUBE with its 24 loudspeakers has the largest playback system (approx. 120 m^2 surface area) available at the Institute for Acoustics and Electronic Music (IEM) and is used for lectures, as well as concerts, workshops and rehearsals.
For the evaluation of the surround sound setup, a total of 720 Ambisonic room impulse responses from 30 microphone positions to 24 Tannoy[1] System 1200 loudspeakers with two Soundfield ST450 MKII microphone arrays were available for assessment[2]. The microphone positions were distributed in a 1-by-1 m spaced grid and the loudspeaker positions were provided in the data set. In fig. 4.2, the measurement grid of the microphone arrays and positions of the loudspeakers are illustrated.

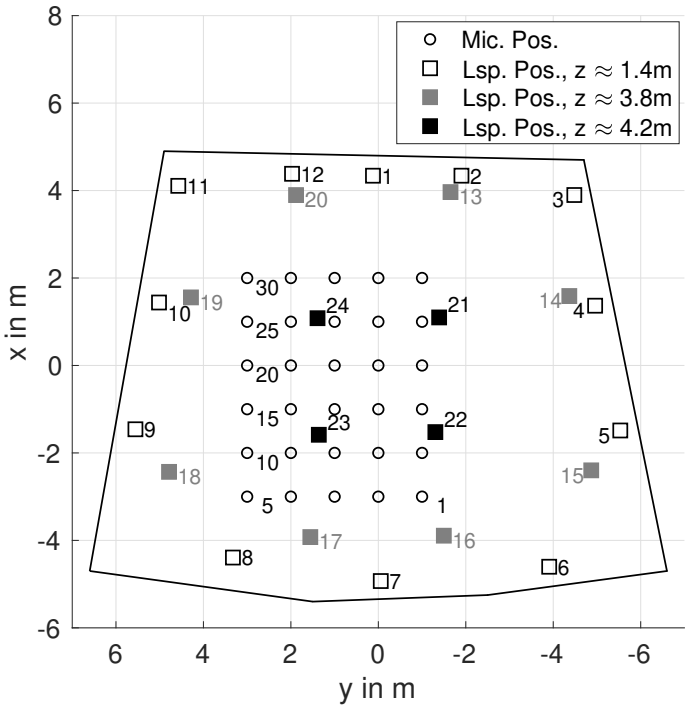


Figure 4.2.: Microphone and LSP positions of the measured set of ARIRs (see [50]).

4.2.1. Localization Accuracy

The accuracy of the estimated directions and positions is evaluated by comparing them to the loudspeaker positions which were assessed by measuring physical distances in the room. The estimation of the loudspeaker directions serves as foundation for the subsequent prediction of the phantom source directions. In the following sec sec. 4.2.1.1, it is investigated, how well the estimation method is performing at a center and an off-center position.
In order to make valid statements about the influence of loudspeaker direction mismatches, the phantom source positions are predicted with the extended energy vector model, based on the uncorrected and corrected loudspeaker estimations from the central microphone position (see sec. 4.2.1.2).

[1] https://www.tannoy.com
[2] The measurements were conducted by Kaspar Müller in the context of his Master's thesis and are available under a public license at https://phaidra.kug.ac.at/view/o:104435 [50]

4.2.1.1. Directions of Loudspeakers

Although the author recommends the estimation of loudspeaker directions from a central microphone position, the assessment at the outer microphone position 5 is conducted in order to investigate the robustness of the implemented estimation algorithm.

Accuracy at the Central Microphone Position The central microphone position in the IEM CUBE is located at position 17 ($[x = 0, y = 0, z = 1.4]$, see fig. 4.2). The influence of the azimuthal error is depending on the zenithal position: An azimuthal error of any size at the zenith is practically non-existent

$$\varphi \sin(\vartheta) = \varphi \sin(0°) = 0\,° \quad \forall \quad \varphi = [-180°, 180°], \tag{4.2}$$

whereas an error of $1\,°$ in the horizontal plane ($\vartheta = 90\,°$) is fully exhibited

$$\varphi \sin(\vartheta) = 1° \sin(90\,°) = 1\,°\,. \tag{4.3}$$

Therefore, the angular deviation e_{total} between the measured and estimated direction vector is used as an overall error metric:

$$e_{\text{total},i} = \arccos(\langle \boldsymbol{l}_{i,\text{meas}}, \boldsymbol{l}_{i,\text{est}} \rangle)\,. \tag{4.4}$$

For the calculation of the scalar product of $\boldsymbol{l}_{i,\text{meas}}$ and $\boldsymbol{l}_{i,\text{est}}$, both vectors are converted into Cartesian vectors of length 1.
The direction estimation error in the center has the median $\tilde{e}_{\text{total}} = 7.19\,°$ and is spread over the interval $[3.43\,°, 11.48\,°]$. The maximum of the azimuthal error in the horizontal plane is $10.91\,°$ at loudspeaker 9.
Analyzing the estimated zenithal angles, a substantial downward drift, detectable through the calculation of the mean error in the zenith $\bar{e}_\vartheta = 5.41\,°$) at almost all positions can be noticed (see fig. 4.3). The suspicion of an inclination of the microphone can be excluded, since the estimation would drift in one half downwards, but in the opposite half upwards. As well, a wrongfully determined microphone height is not a plausible reason, since the position had to be raised by more than 40 cm in order to compensate this zenithal error. A possible explanation could give the design of the microphone; due to its physical limits, e. g., small deviations in the manufacturing and unbreachable time delays between the membranes, it can never behave like an ideal, coincident array. Further investigations could check if the drift changes in its opposite direction, when the orientation of the microphone is turned upside down. Fig. 4.3 shows a projection of true (measured) and estimated loudspeaker directions on a plane. This graphical presentation is helpful to get a first overview, but one should always be aware of the distortion of azimuthal angles towards the zenith and nadir due to the projection.

Figure 4.3.: Measured and estimated LSP directions from the central mic. pos. 17.

Bias Corrections in Practice One big advantage of a successful loudspeaker direction estimation is the possibility to avoid time-consuming position measurements by hand. One way to improve the estimation accuracy is to determine the position of a single loudspeaker and align all estimations relative to its direction. The total median error can be decreased to $\tilde{e}_{\text{total}} = 6.49\,^\circ$ with an alignment to the measured position of loudspeaker 1. The zenithal error decreases, whereas the azimuthal error in the horizontal axis increases, since the estimation of loudspeaker 1 was also shifted in the azimuthal direction. The potential of the bias correction can only be as successful as the estimation accuracy of the loudspeaker, which is chosen for the alignment, allows it.

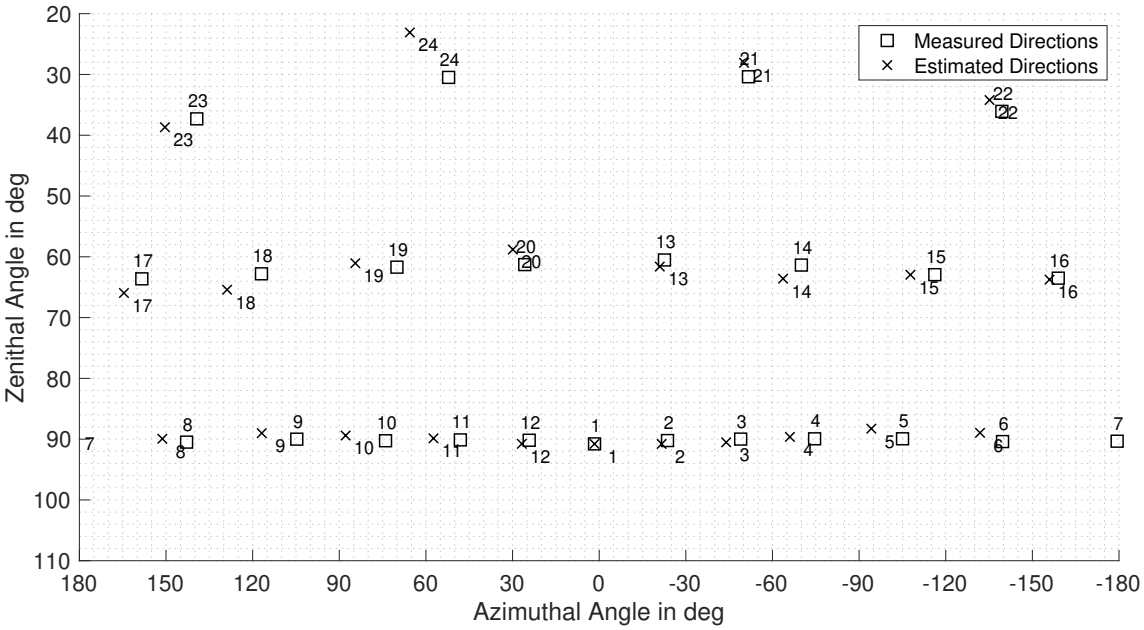


Figure 4.4.: Measured and bias-corrected, estimated LSP directions from the central mic. pos. 17.

Tab. 4.1 enables a quick comparison of the error values. The direction estimation in the horizontal plane is important in the context of the error estimation of the lateral energy fraction in sec. 3.3. Fig. 4.4 shows the estimations after the bias correction.

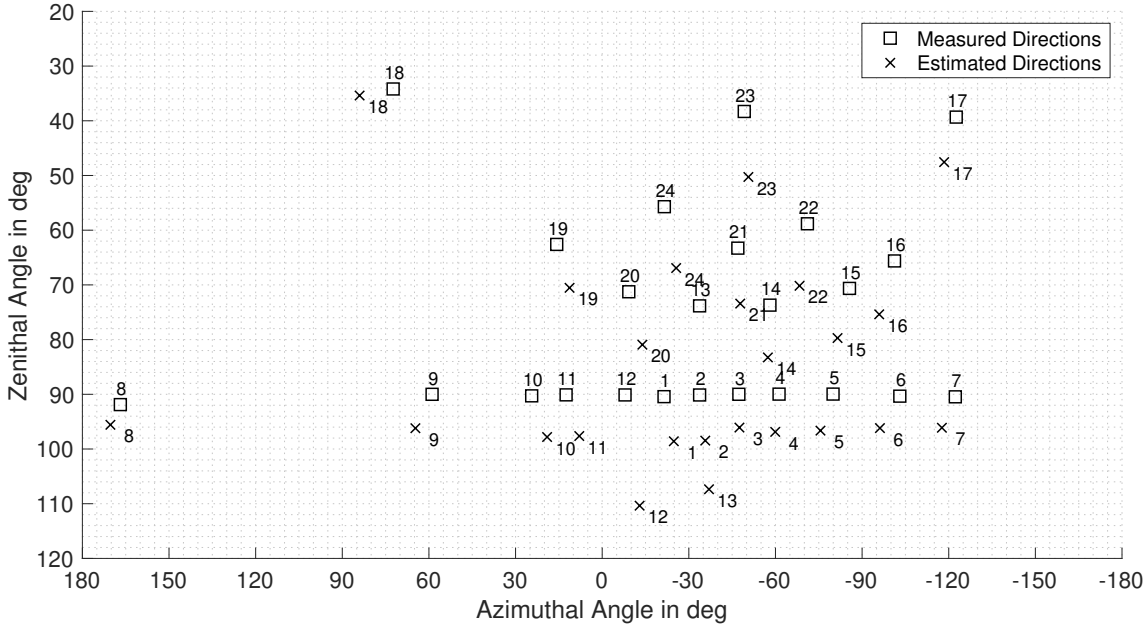
	Biased in °	Corrected in °
Median Error $\tilde{e}_{\text{total}}$	7.19 °	6.49 °
Median Azim. Error $\tilde{e}_{\varphi}$ of LSPs 1-12	4.59 °	8.10 °
Median Zen. Error $\tilde{e}_{\vartheta}$	5.58 °	1.04 °

Table 4.1.: Results of the DOA estimation from the center mic. pos. 17.

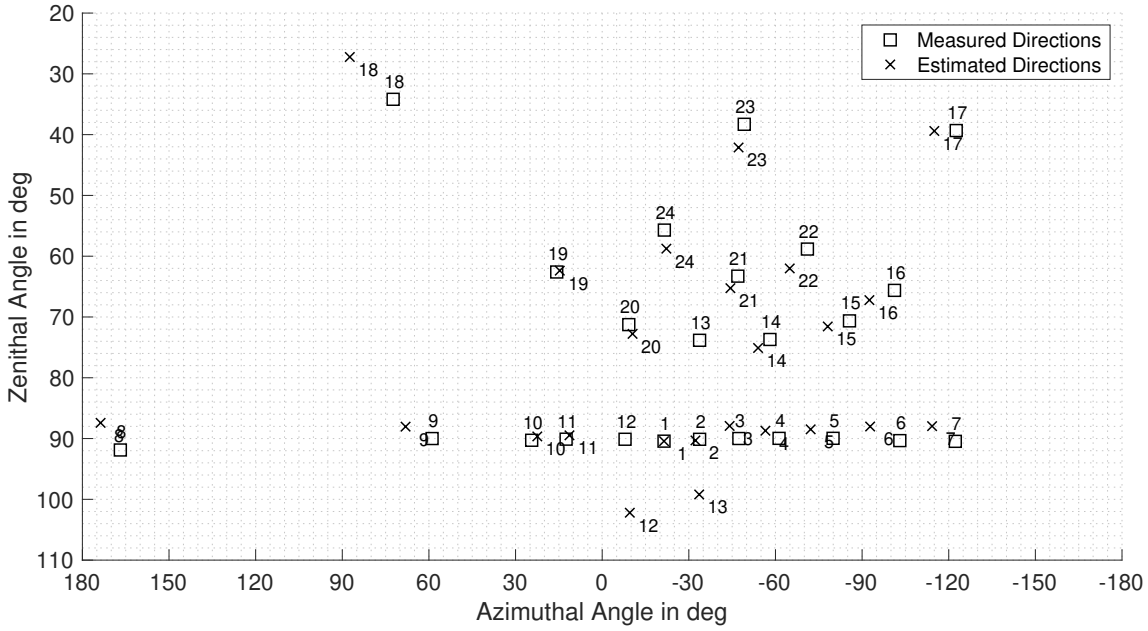
Accuracy at an Off-Center Microphone Position In order to investigate the influence of the proximity to the walls of the room and the thereby accompanying stronger early reflections, the direction estimation from microphone position 5 is analyzed.
Again, a downwards drift deteriorates the measurement. The large errors of loudspeaker 12 and 13 can be explained by looking at the course of the particular PIV $\boldsymbol{I}(t)$. Fig. 4.6 shows the PIV of the impulse response of loudspeaker 12 as an example. The algorithm interprets the maximum of the pseudo-intensity as the arrival of the first wave front. Under specific spatial constellations, the first reflection can be energetically more intense at the microphone position than the first wave front. This mismatch can easily be corrected by searching for the peak of the pseudo-intensity vector and detect possible earlier peaks over a threshold T, which is defined relative to the detected maximum value, e. g., $T = 0.9 \max\{\boldsymbol{I}(t)\}$. This phenomenon occurs only at off-central positions in this set of measurements.
The bias correction improves all error measures (see tab. 4.2), whereas the mapping errors for loudspeaker 12 and 13 stay preserved.

	Biased in °	Corrected in °
Median Error $\tilde{e}_{\text{total}}$	8.96 °	4.85 °
Median Azim. Error $\tilde{e}_{\varphi}$ of LSPs 1-12	4.46 °	3.96 °
Median Zen. Error $\tilde{e}_{\vartheta}$	8.19 °	1.78 °

Table 4.2.: Results of the DOA estimation from the off-center mic. pos. 5.



(a) Without bias correction.



(b) With bias correction.

Figure 4.5.: Measured and estimated LSP directions from mic. pos. 5.

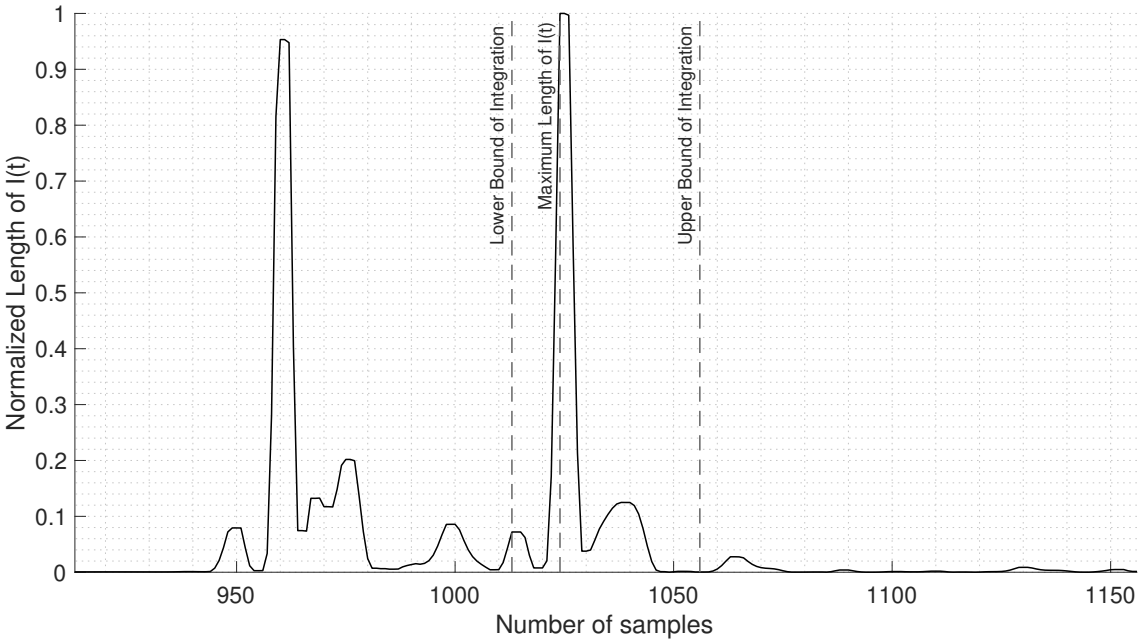
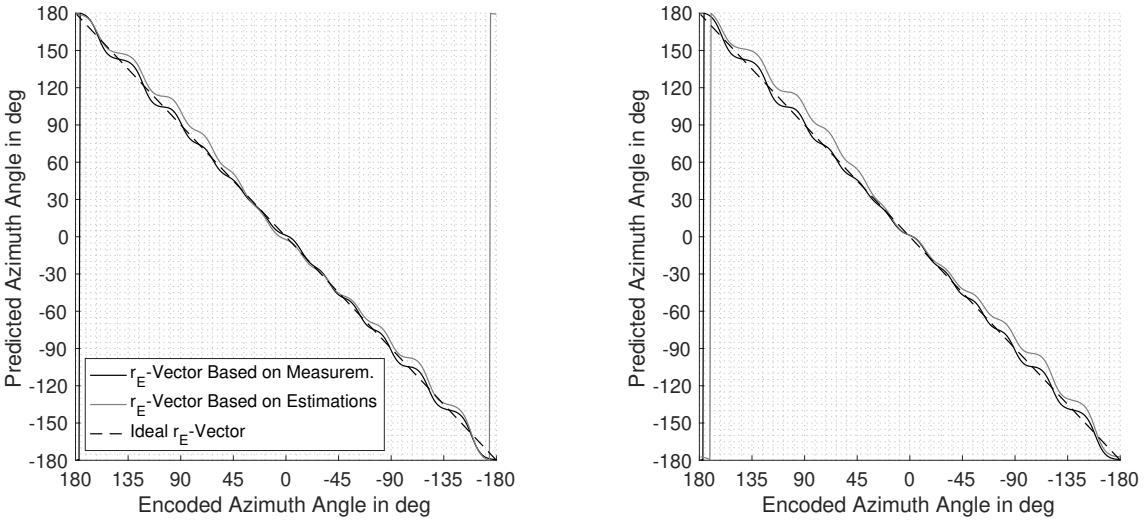


Figure 4.6.: Course of the normalized length of the PIV $\boldsymbol{I}(t)$ of the impulse response from LSP 12 measured at mic. pos. 5.

4.2.1.2. Directions of Phantom Sources

With the results of the loudspeaker direction estimation and the knowledge of the loudspeaker gains, the localization of phantom source positions can be predicted with the extended $\boldsymbol{r}_{\mathrm{E}}$-vector model (see sec. 2.3).
In this section, the influence of loudspeaker mismatches on the prediction of the phantom source direction is investigated. The energy vector on basis of the measured loudspeaker positions is calculated and defined as true reference (solid black line) and compared to the results calculated with the estimated directions (solid grey line). Additionally, the ideal course of the phantom source according to the encoded directions is marked in the presented figures with a dashed, black line.
The movements are assessed on two trajectories: The first one leads in a full circle around the horizontal plane ($\varphi = [\text{-}180°, 180°], \vartheta = 90°]$) and the second from the nadir over the zenith back to the nadir ($\varphi = 0°, \vartheta = [\text{-}180°, 180°]$). As phantom source signal, a Dirac impulse $\delta(t)$ of height 1 is encoded in 7th-order Ambisonics and decoded to the desired position with the basic-weighted AllRAD-decoder.

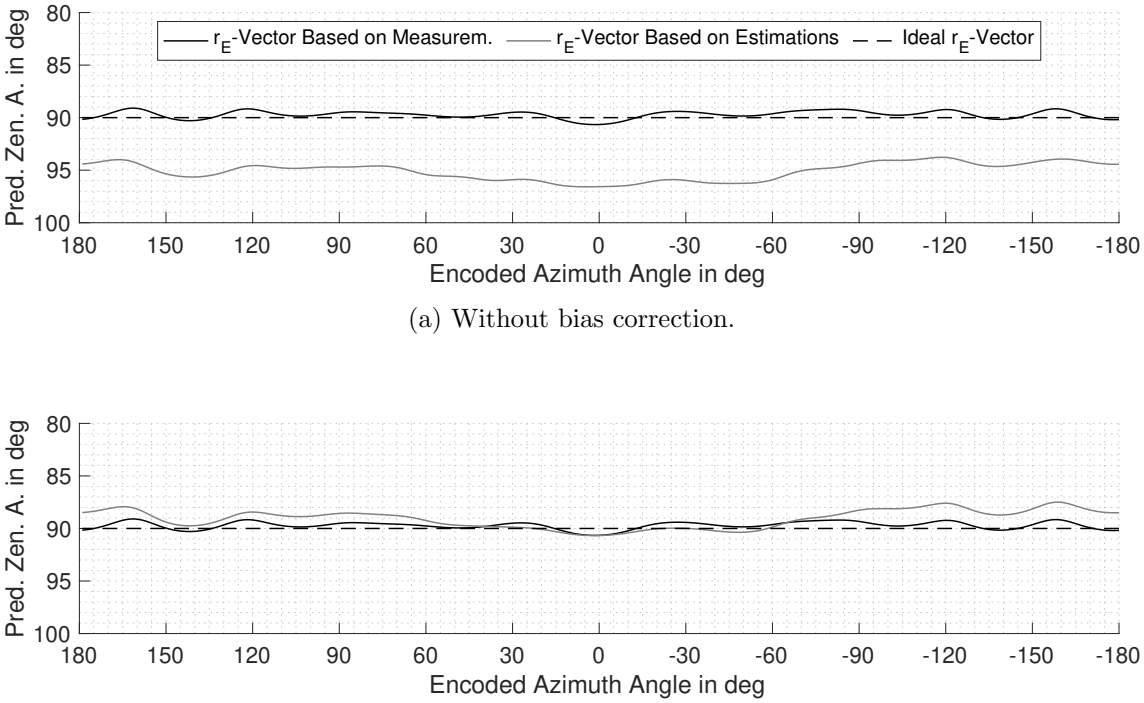
In the Horizontal Plane The localization prediction of a phantom source moving 360 ° around the horizon is investigated first. The prediction is very similar to the correct prediction but the deviation increase while moving away from the front. The median azimuthal error equals $\tilde{e}_{\varphi} = 4.32\,°$. The zenithal bias towards the nadir propagates from the loudspeaker localization and is clearly recognizable in fig. 4.8a. It yields an offset in the localization prediction. In the zenithal dimension, the median error equals $\tilde{e}_{\vartheta} = 5.37\,°$.
Comparable to the bias correction of the loudspeaker direction, the prediction of the zenithal localization is improved on the cost of a worsened azimuthal localization (see tab. 4.3).



(a) Without bias correction.

(b) With bias correction.

Figure 4.7.: Prediction of the perceived azimuth angle φ of the moving phantom source.

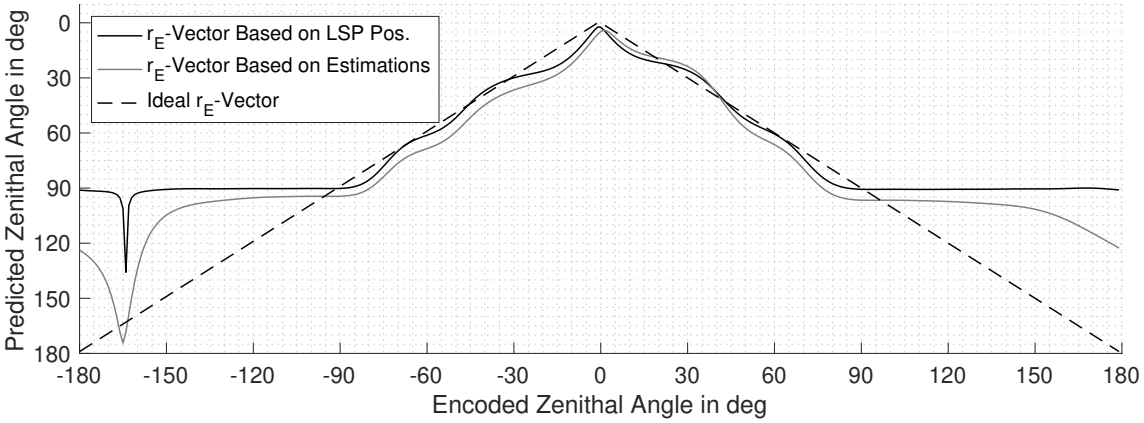


(a) Without bias correction.

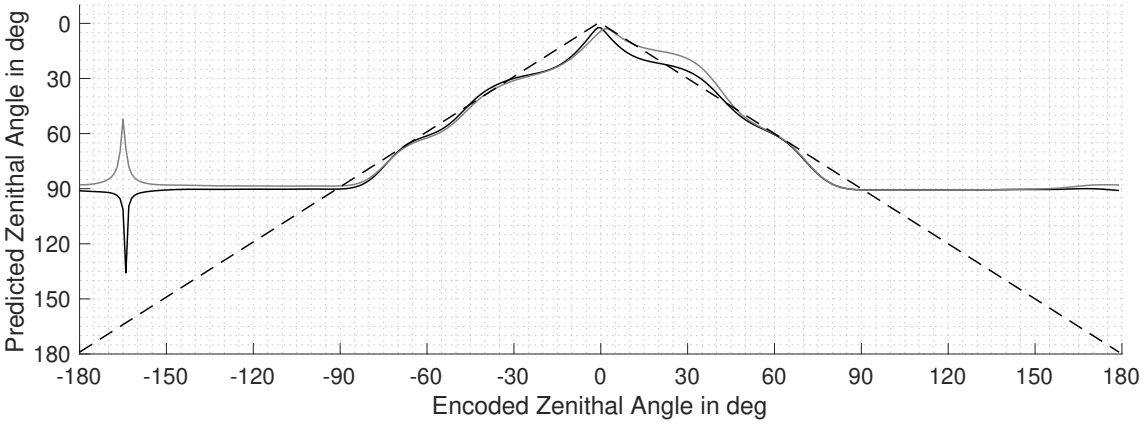
(b) With bias correction.

Figure 4.8.: Prediction of the perceived zenith angle ϑ of the moving phantom source.

In the Median Plane The prediction of the phantom source position works accurately with a median error of $e_\vartheta = 5.95\,^\circ$ below the horizontal plane. The alignment of the estimation to loudspeaker 1 minimizes the median zenithal error to $\tilde{e}_\vartheta = 1.54$ (see tab. 4.3). Analyzing the encoded zenithal directions in fig. 4.9, one can observe that there are only ϑ-values $\leq 90\,^\circ$ possible, since there are no loudspeakers for $\vartheta > 90°$. For the notches in both curves around -175 °, no plausible explanation could be found.



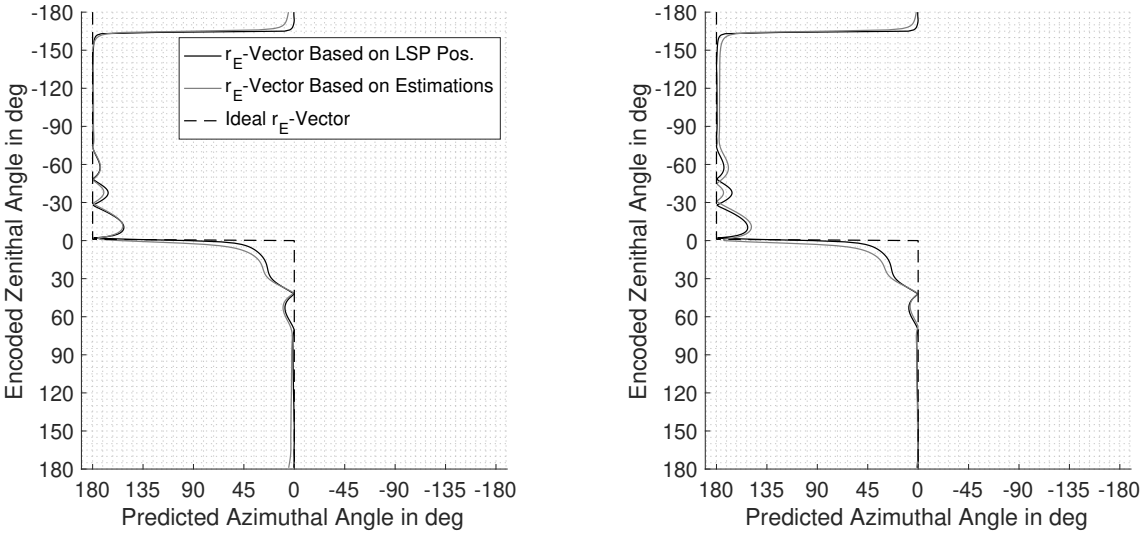
(a) Without bias correction.



(b) With bias correction.

Figure 4.9.: Prediction of the perceived zenith angle ϑ of the moving phantom source.

Fig. 4.10 shows the predicted azimuth angle φ based on the uncorrected (fig. 4.10a) and corrected loudspeaker position estimations (fig. 4.10b) for the vertical trajectory from the nadir to the zenith and back. All deviations from the ideal prediction happen in the upper 60 ° of the zenith angle ϑ.



(a) Without bias correction.

(b) With bias correction.

Figure 4.10.: Prediction of the perceived azimuth angle φ of the moving phantom source.

		Movement in the Horizontal Plane	Movement in the Median Plane
Azim. Error $\tilde{e}_\varphi$ in °	biased	4.32	-
	corrected	7.82	-
Zen. Error $\tilde{e}_\vartheta$ in °	biased	5.37	5.95
	corrected	0.67	1.54

Table 4.3.: Overview over the biased and corrected median of azimuthal and zenithal errors $\tilde{e}_\varphi$ and $\tilde{e}_\vartheta$.

4.2.1.3. Conclusion on the Localization Performance

The localization performance of the presented method proved to be accurate and robust enough for the further use as foundations for the localization prediction of phantom sources. Through a bias correction of the loudspeaker direction estimation, the total angular mismatch could be further reduced. Obviously, mapping errors in the DOA estimation of loudspeakers translate directly into the prediction with the $\boldsymbol{r}_\mathrm{E}$-vector model. The detected ripples, where the curves of both energy vector model computations deviate from the ideal course can be explained through the not ideal and sparse placement of the loudspeakers in the elevated rings.
Overall, the requirements of the thresholds (see sec. 2.1) regarding the median errors could be met. The uncorrected median error in the azimuth falls below the defined 5 °-threshold and the median zenithal error in both trajectories reclines under the 10 °-threshold, in the corrected scenario, $\tilde{e}_\vartheta$ is even smaller than 2 °.

4.2.2. Adapted Direct-to-Reverberant Energy Ratio (DRR_m)

In this section, the influence of the two spatialization methods AllRAD and VBAP on the direct-to-reverberant energy ratio is studied. For the decoding and panning of the phantom source, the measured loudspeaker positions are used. For all following evaluations of the DRR_m in this section, an ideal Dirac impulse $\delta(t)$ of magnitude $\delta(0) = 1$ is reproduced as discrete or phantom sound source. The figures show the contour plots of the DRR_m in dB for the area in the IEM CUBE which is covered by microphone positions. The positions are marked with grey circles and for a higher resolution in the plot, an interpolation between the calculated values at the microphone positions was conducted.

Discrete Sound Source Fig. 4.11 shows the DRR_m distribution, when loudspeaker 1 ($\boldsymbol{\theta}_1 = [1.61°, 90°]$) relative to the central microphone position) is active. As expected, the near field of the loudspeaker, the DRR_m exhibits high values that fall with increasing distance. At the left bound of the investigated area, the diffuse sound energy is equal to the energy of the direct sound (0 dB).

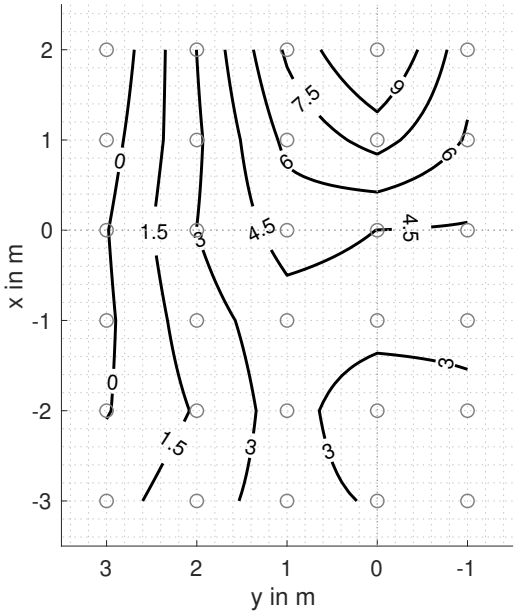
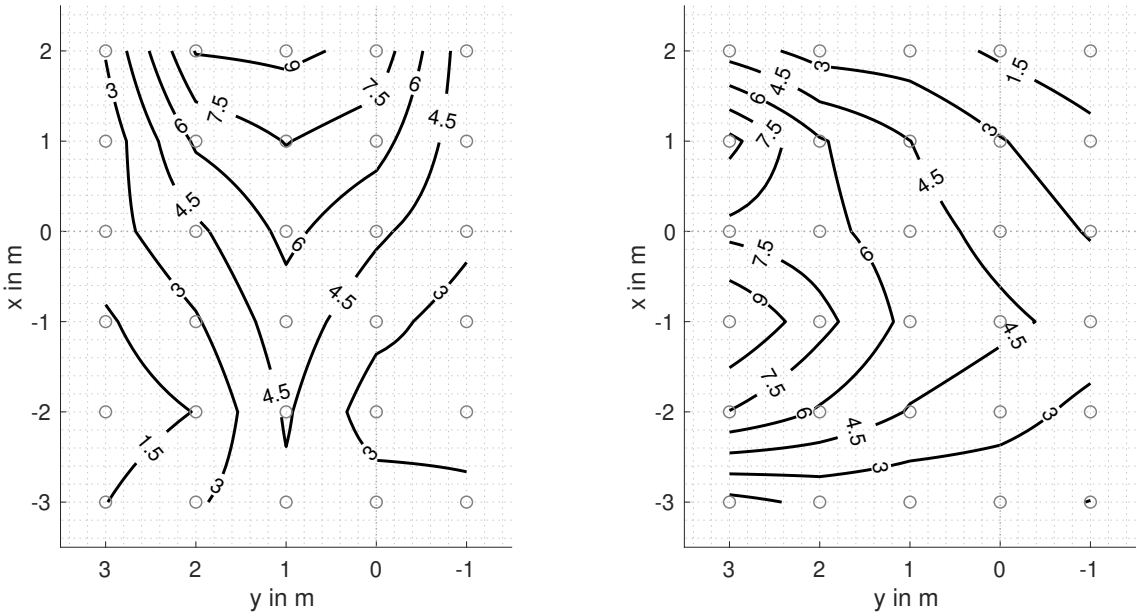


Figure 4.11.: DRR_m contour plot in dB of a discrete sound source at LSP 1.

Phantom Sources Using VBAP Encoding a phantom source at the position of loudspeaker 1 yields the same results as above in fig. 4.11, since only loudspeaker 1 will be active. Fig. 4.12a shows a phantom source encoded between speaker 1 and 12 ($\boldsymbol{\theta} = [12.93°, 90°]$). It is clearly recognizable that now the area of the source is widened and the DRR_m decreases slower with increasing distance. Furthermore, the 0 dB limit is now shifted outwards of the assessed area. When panning a source to the left side ($\boldsymbol{\theta}_{\text{phsrc}} = [90°, 90°]$), even the two separate beams of the two active loudspeakers are distinguishable in the plot (s. fig. 4.12b). The panning direction lies in the middle between two loudspeakers and since they are more distant to each other than two frontal loudspeakers, the two beams become distinguishable.

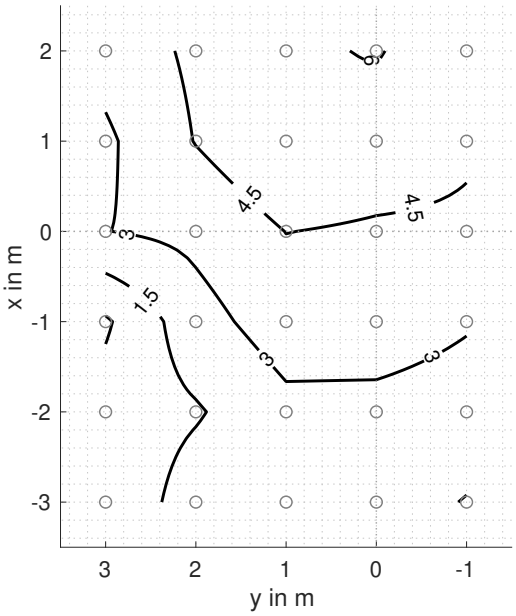


(a) Panning direction between LSP 12 and 1 ($\boldsymbol{\theta} = [12.93°, 90°]$)).

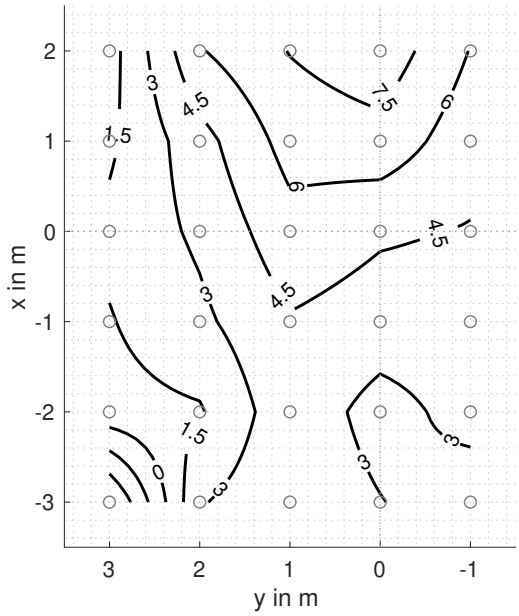
(b) Panning direction between LSP 9 and 10 ($\boldsymbol{\theta} = [90°, 90°]$)).

Figure 4.12.: DRR_m contour plots in dB of a phantom source panned with VBAP.

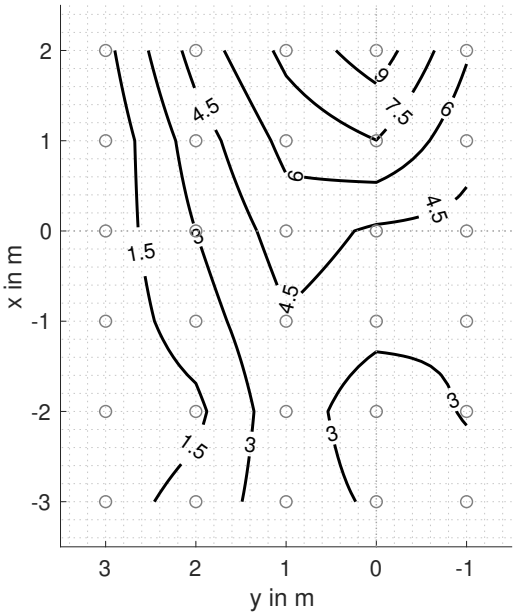
Phantom Sources Using Ambisonics The performance of Ambisonics depends highly on the chosen order, the decoding approach and the used decoding weights. For all following plots, the AllRAD decoder (presented in sec. 1.2.3.2) and basic weighting was used.
From fig. 4.13a to fig. 4.13d the phantom source is en- and decoded with increasing Ambisonic order $N = 1,\ 3,\ 5,\ 7$ in the direction of loudspeaker 1 ($\boldsymbol{\theta} = [1.61°, 90°]$)). The decline of DRR_m is shifted in the negative x-direction and the direct sound area gets narrower. The DRR_m distribution converges with increasing order to the distribution of the discrete sound source. Except some minor discrepancies, the contour plot of the phantom source, encoded in 7th-order Ambisonics, equals fig. 4.11. This trend is also perceivable by looking at the loudspeaker gains. The higher the order, the less other loudspeakers besides loudspeaker 1 are active. In fig. 4.14, the gains of 1st-order (FOA), 3rd-order (3OA), 5th-order (5OA), 7th-order (7OA) Ambisonics, and VBAP (corresponding to the discrete source) for the phantom source reproduction at loudspeaker position 1 are presented.
Apparently, the use of multiple loudspeakers augments the sweet area ($DRR \geq 0\,\text{dB}$). Comparing fig. 4.13c and 4.13d to the direct source scenario in fig. 4.11, the contours in the sweet area stay almost unchanged, while the 0 dB limit is pushed to the left.



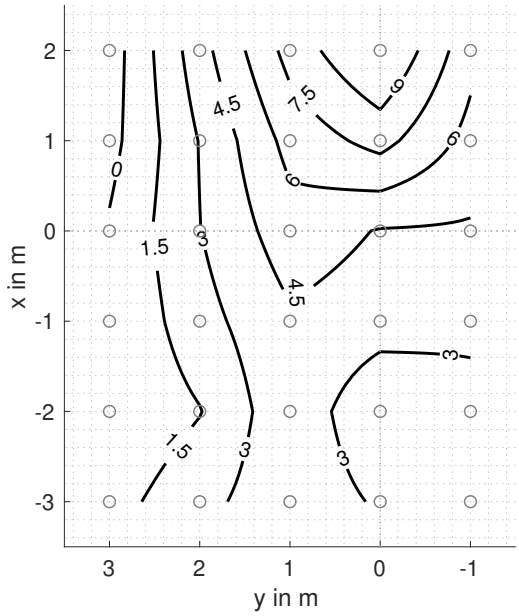
(a) 1st-order AllRAD decoding.



(b) 3rd-order AllRAD decoding.



(c) 5th-order AllRAD decoding.



(d) 7th-order AllRAD decoding.

Figure 4.13.: DRR_m contour plots in dB of phantom sources decoded with basic-weighted Ambisonics.

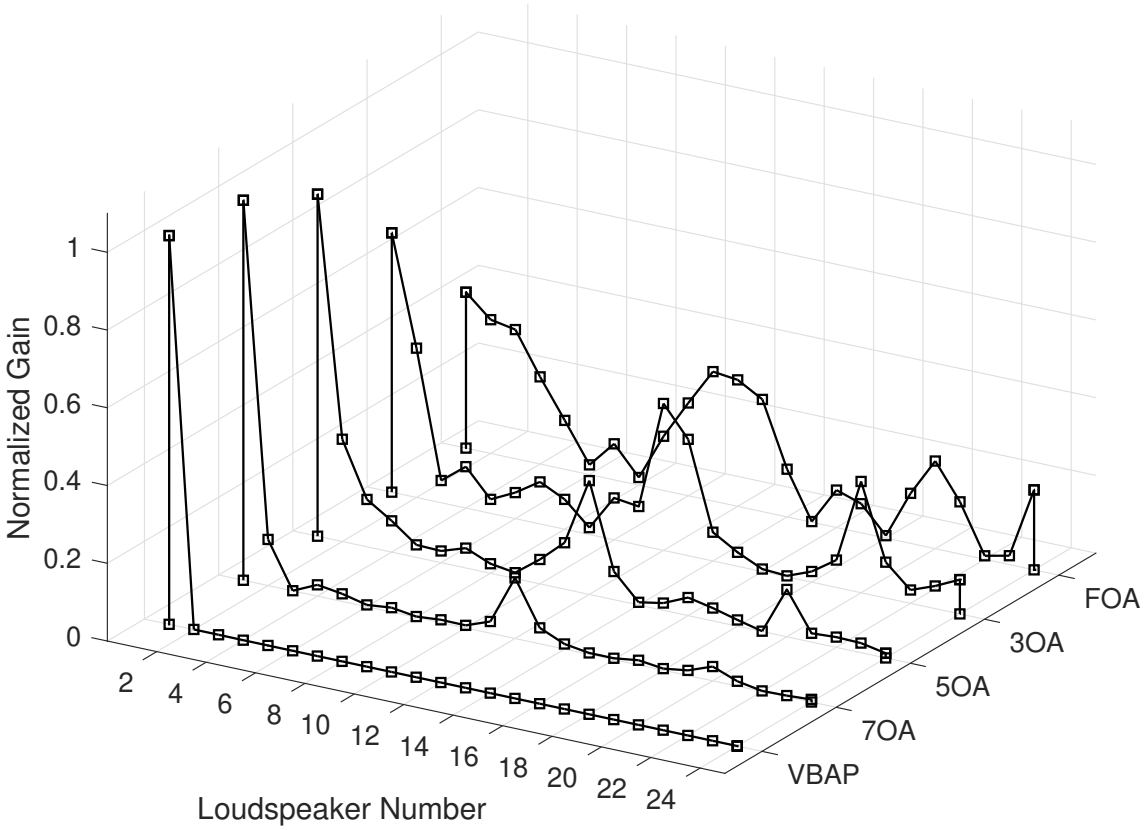


Figure 4.14.: Normalized loudspeaker gains for all 24 LSPs using different spatialization methods.

4.2.3. Adapted Lateral Energy Fraction (LF_m)

For the evaluation of the apparent source width (ASW) using the LF_m in the horizontal plane ($\vartheta = 90\,°$), a Dirac impulse is decoded via AllRAD or panned via VBAP in every azimuthal orientation with a resolution Δ_φ of 1° and subsequently analyzed at the central microphone position 17. The LF_m values for the discrete sources of singular loudspeakers are marked with squares in the plots.

Phantom Sources Using VBAP When using VBAP for panning phantom sources in the horizontal plane, at most two loudspeakers are active. As described in sec. 1.1.2, when positioning a phantom source in the exact direction of a loudspeaker, only this particular loudspeaker will be active. Thus, it is obvious that the LF_m values in the direction of the loudspeakers coincide exactly with the values of the assessment of discrete sound sources. Between two loudspeakers, the LF_m increases, which agrees with the larger ASW described in the results of Frank's listening experiment presented in [45, p. 48f].
Surprisingly, the values of the discrete sound sources vary. Especially the ones of the loudspeakers 1, 3, 7, and 11 are significantly lower than their neighbors. Through the spatial conditions of the loudspeaker positions, the lateral reflections at the microphone position of these auditory events are very sparse: LSP 1 and 7 are located in the middle of the front and rear wall at the most distant point to the side walls. The setup of the loudspeakers 3 and 11 in the frontal left and right corners does not yield strong lateral reflections as well.

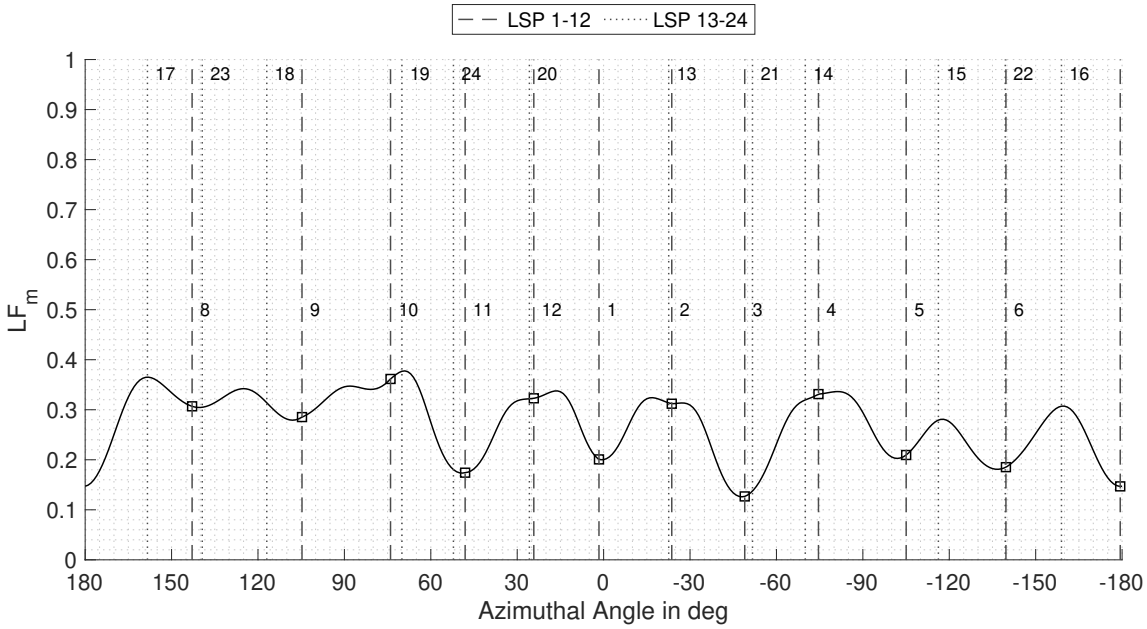
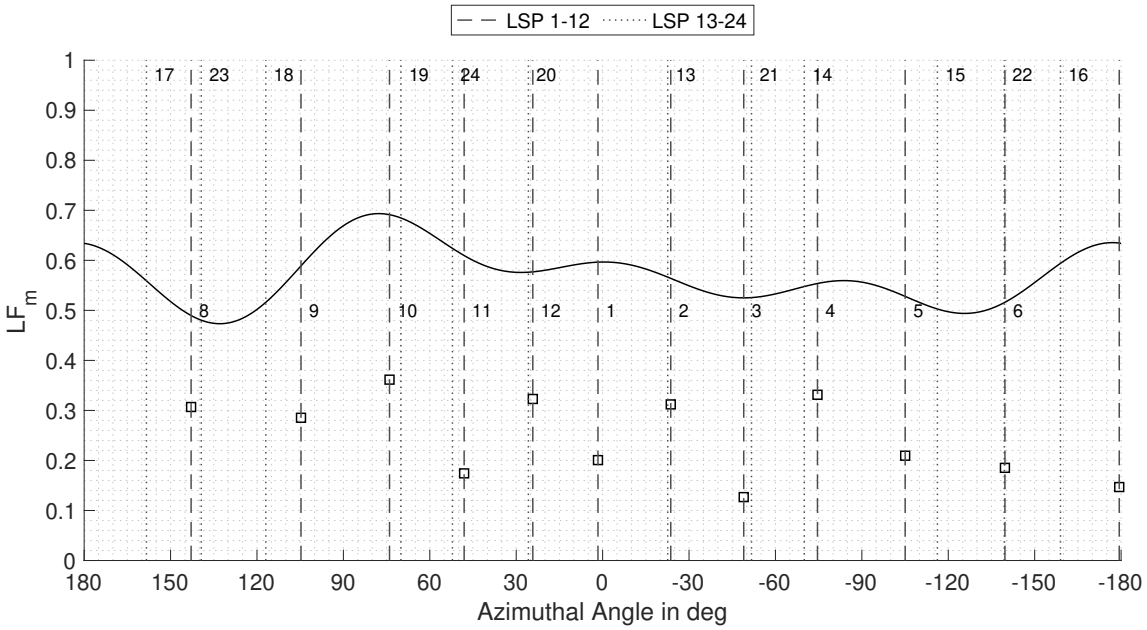
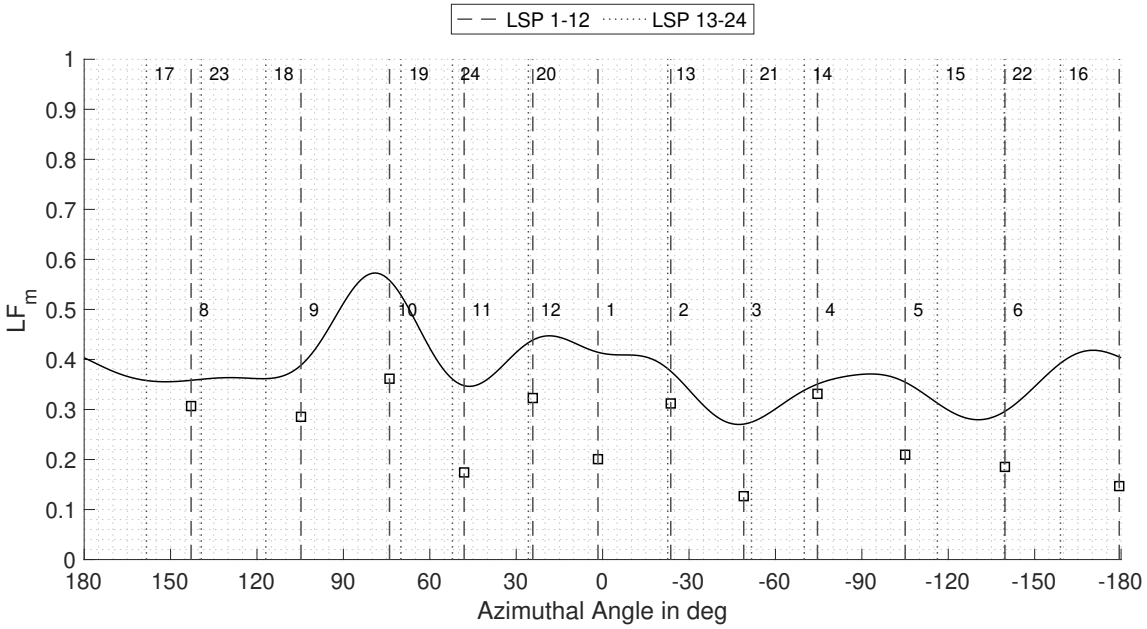


Figure 4.15.: LF_m for a phantom source in the horizontal plane using VBAP.

Phantom Sources Using Ambisonics This paragraph shows the different LF_m-results for AllRAD decoding using basic weighting in different orders.
The 1st Ambisonic order (see fig. 4.16a) produces a very wide and more consistent ASW with lesser drops than VBAP for the price of a much higher source spread. The perceived source width is increased, since more loudspeakers participate at the synthesis of the phantom source.
The same trend as in the previous section 4.2.2 is observable: Fig. 4.16b and fig. 4.17b show that the higher the Ambisonic order is, the more the gains and the ASW approximates to the results of VBAP. The remaining deviations from the VBAP result in 7th-order Ambisonic are caused by the stronger participation of the elevated loudspeakers, which remain inactive when VBAP is used.

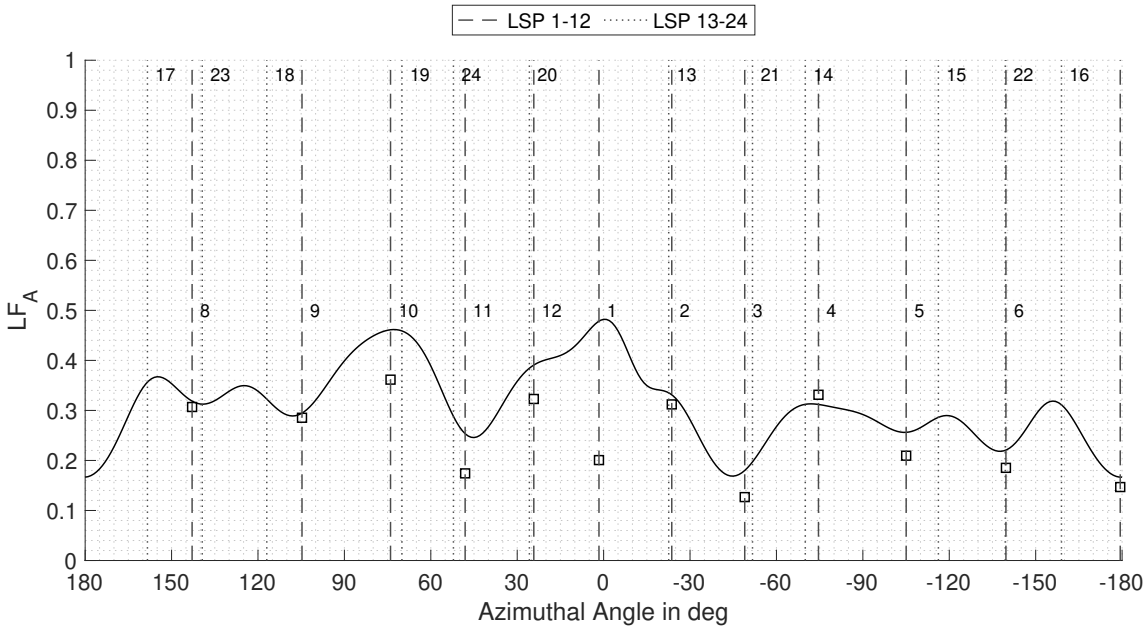


(a) 1st-order Ambisonics.

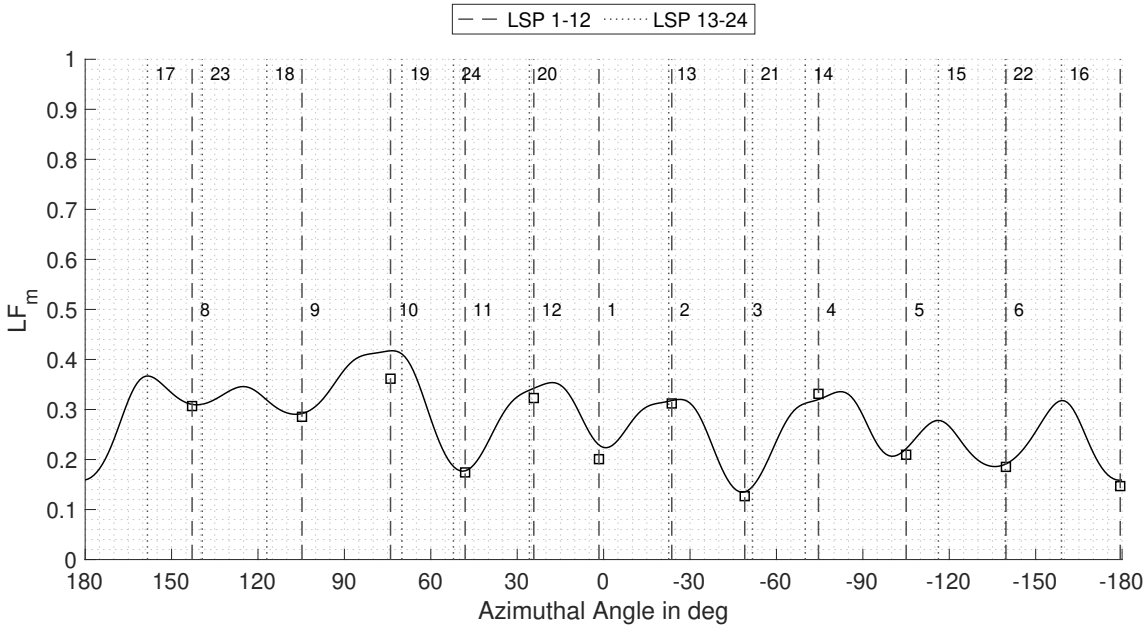


(b) 3rd-order Ambisonics.

Figure 4.16.: LF_m for a phantom source on a circle in the horizontal plane using AllRAD and basic weighting.



(a) 5th-order Ambisonic.



(b) 7th-order Ambisonic.

Figure 4.17.: LF_m for a phantom source on a circle in the horizontal plane using AllRAD and basic weighting.

4.2.4. Differences in Coloration (C_m)

According to Frank's procedure presented in sec. 3.4, a phantom source is panned clockwise in six steps of approx. 5 ° from loudspeaker 1 ($\boldsymbol{\theta}_1 = [1.61°, 90°]$) to loudspeaker 2 ($\boldsymbol{\theta}_2 = [-23.65°, 89.52°]$) and the spectral differences to the left neighboring position are calculated and plotted. The differences in coloration are again evaluated at microphone position 17.

Phantom Sources Using VBAP The difference plots in fig. 4.18 exhibit several strong extrema with a magnitude over 2 dB. The maximum $C_m = 2.5$ dB can be found at 100 Hz.

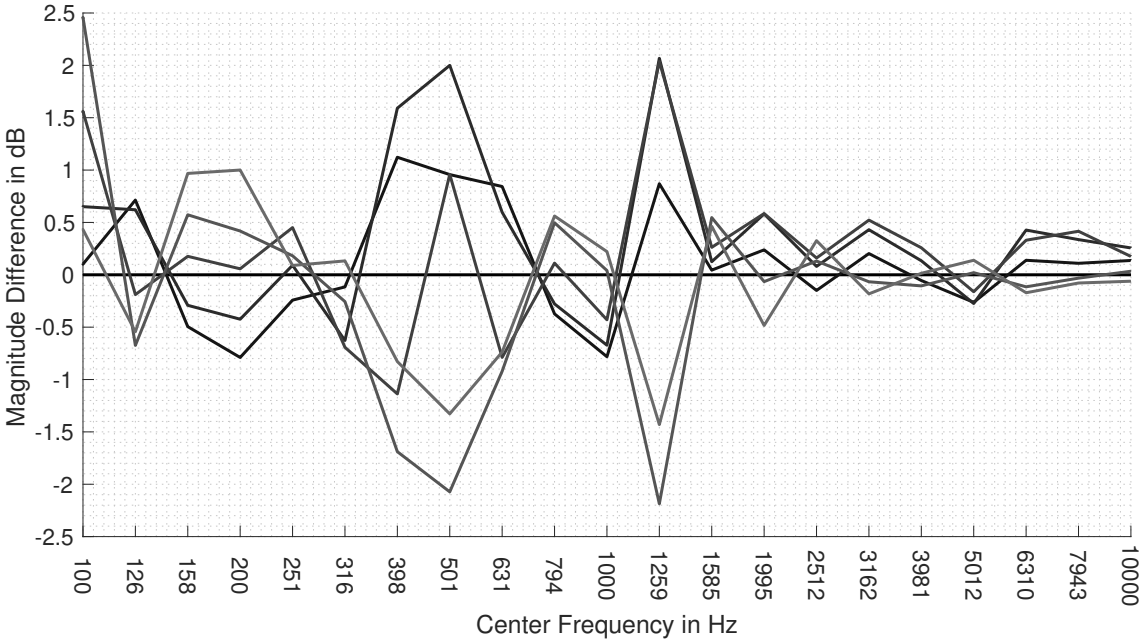
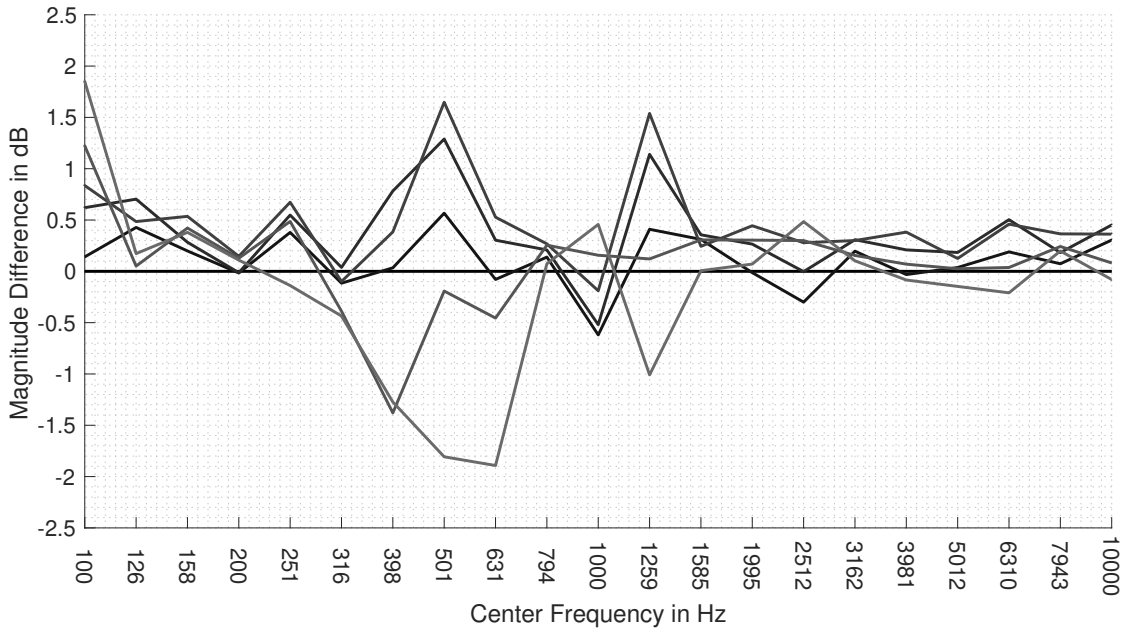


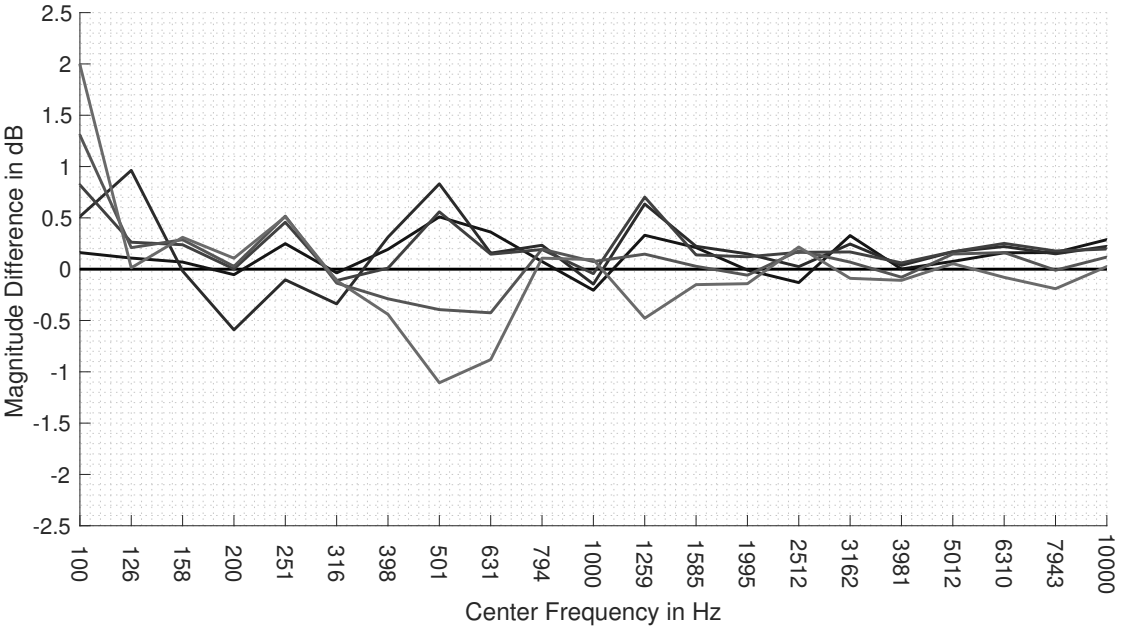
Figure 4.18.: Spectral difference plots in third-octave bands using VBAP.

Phantom Sources Using Ambisonics The coloration is strongly depending on the used Ambisonic order N. The higher the order, the bigger the differences in coloration. Using order $N = 7$ and basic weighting produces very VBAP-like peaks and notches in the same frequency range (see fig. 4.19a). However, the differences are not as large: All extrema lie under 2 dB. The largest difference is $C_m(631\,\text{Hz})$ =-1.89 dB.
As already presented by Frank in [45, p. 84], the differences in the mid frequencies can be further attenuated by an adaption from basic to max-$\boldsymbol{r}_\text{E}$ weighting. Fig. 4.19a shows that the maximum rises from 1.85 dB to 1.99 dB at 100 Hz, but almost all peaks and notches in the mid-frequency range are observably reduced under 1 dB absolute difference and are therefore barely perceivable anymore [61]. The new extremum $C_m(501\,\text{Hz})$ equals now -1.11 dB.
Applying max-$\boldsymbol{r}_\text{E}$ weighting in the decoding process yields a reduction of the side lobes but also a broadening of the main lobe (see sec. 1.2.2, par. *Weighting*). Thus, more neighboring loudspeakers (in this scenario especially lsp. 12 and 13) participate at the source reproduction and the strong comb filters are reduced, since more signals superpose each other now. This effect can as well be observed by analyzing the loudspeaker gains in fig. 4.20 for a phantom source direction mid-way between loudspeaker 1 and 2 ($\boldsymbol{\theta} = [-8.49°, 89.81°]$): VBAP uses only two active loudspeakers, whereas the decoding with basic weighting also includes loudspeakers 11, 12, and 13, which already minimizes the strong differences by a

little. The decoding with max-$\boldsymbol{r}_{\mathrm{E}}$ weighting improves the performance further, since the gains of loudspeaker 12 and 13 are increased once more. The lower the Ambisonic order, the lesser the degradation in coloration is exhibited. The difference plots for basic weighted AllRAD-decoding of 1st- and 3rd-order Ambisonics show only not audible fluctuations in the spectrum (see A.4.2).
The presented findings indicate an unavoidable trade-off: In order to achieve a spectrally stable source, the Ambisonic order and thus the spatial resolution has to be reduced or the main lobe of the directivity pattern of the decoding has to be widened, which influences the perceived ASW.



(a) AllRAD and basic weighting.



(b) AllRAD and max-$\boldsymbol{r}_{\mathrm{E}}$ weighting.

Figure 4.19.: Spectral difference plots for third-octave bands using 7th-order.

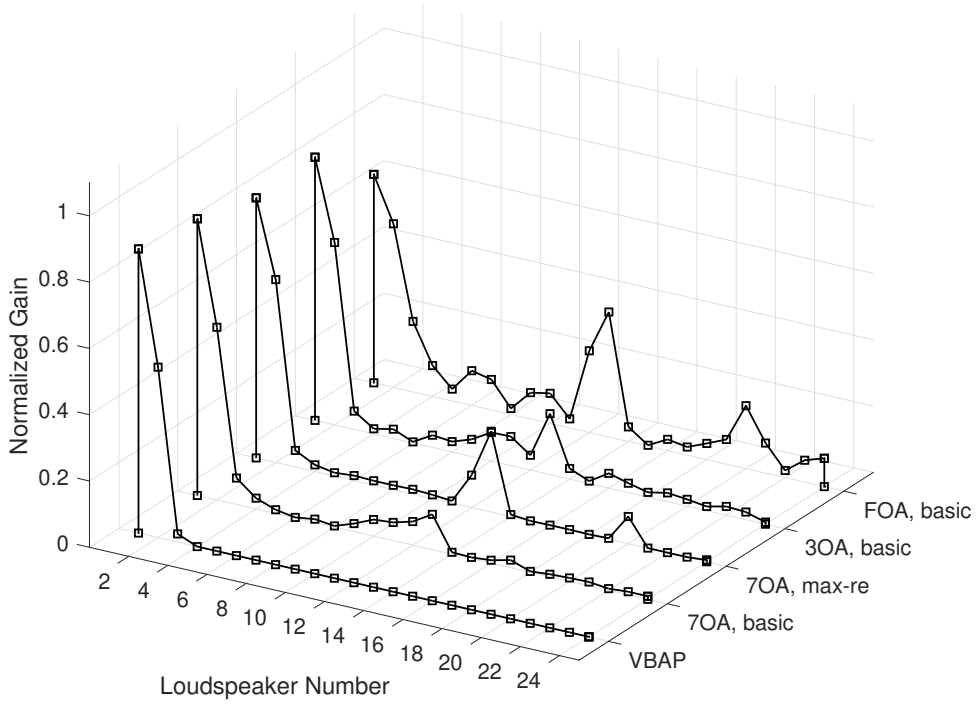
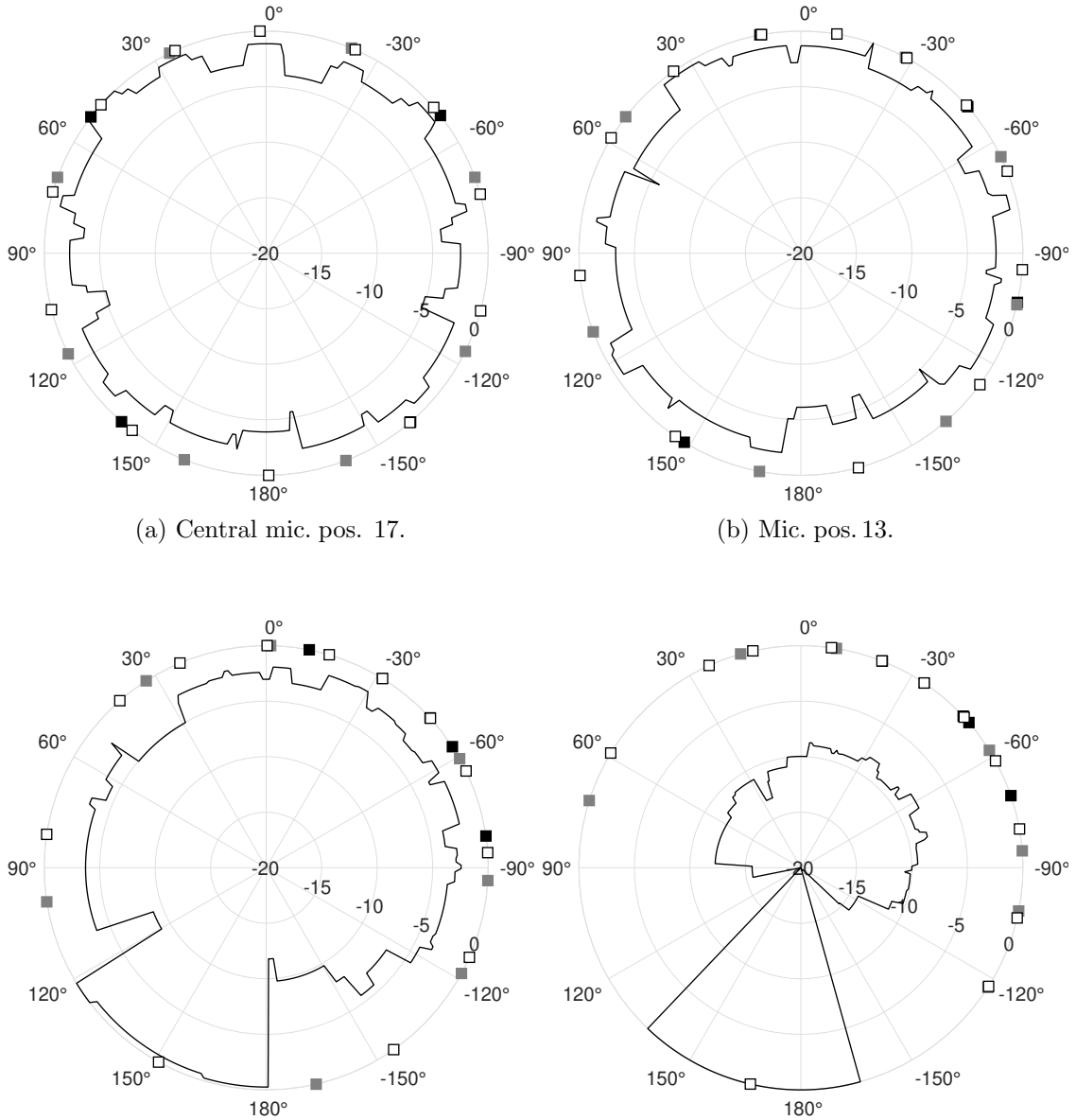


Figure 4.20.: Normalized loudspeaker gains using different spatialization methods and weightings.

4.2.5. Envelopment (EV_m)

For the envelopment assessment, the $EV_m(\varphi)$ is normalized, so that the maximum wedge value equals 0 dB. All values below -20 dB are treated equal regarding their perception. In further calculations and the graphical representations, they are replaced by the threshold value -20 dB. For an easier analysis, the relative directions of the 24 loudspeakers of the audio system are marked with squares at the outer bound of the polar plots (white: lower ring, grey: middle ring, black: upper ring of loudspeakers.).
First, the coverage of the central position, presented in fig. 4.21a, is investigated. The correlation with the loudspeaker directions is recognizable: Spherical wedges containing the most loudspeakers are loudest, whereas wedges with less loudspeakers are quieter. The quietest wedges $\min\{EV_m(\varphi)\}$ are oriented to $\varphi_{\min} = [\text{-}172,^\circ,\text{-}171,^\circ]$.
The standard deviation σ can be used as interpretation of the fluctuation over all spherical wedges. The author assumes that it should be as small as possible for an all-around consistent auditory impression. At the central position, σ is rather small and equals 1.09 dB. Moving diagonally away from the center to the left rear corner of the room, as expected, the level differences between and standard deviations increase: At microphone position 13, the minimum drops to -6.2 dB and the σ rises to 1.24 dB. This trend progresses over position 9 to the outer position 5 (see tab. 4.4).
The closer the microphone is positioned to the corner, the more importance loudspeaker 8 gains. At the last microphone position 5 the spherical wedges containing only loudspeaker 8 are the most prominent and the energy content of the neighboring wedges collapses below -20 dB. So far, the results of the quality measure EV_m support the informal envelopment perception of the author in the IEM CUBE.



(a) Central mic. pos. 17.

(b) Mic. pos. 13.

(c) Mic. pos. 9.

(d) Mic. pos. 5.

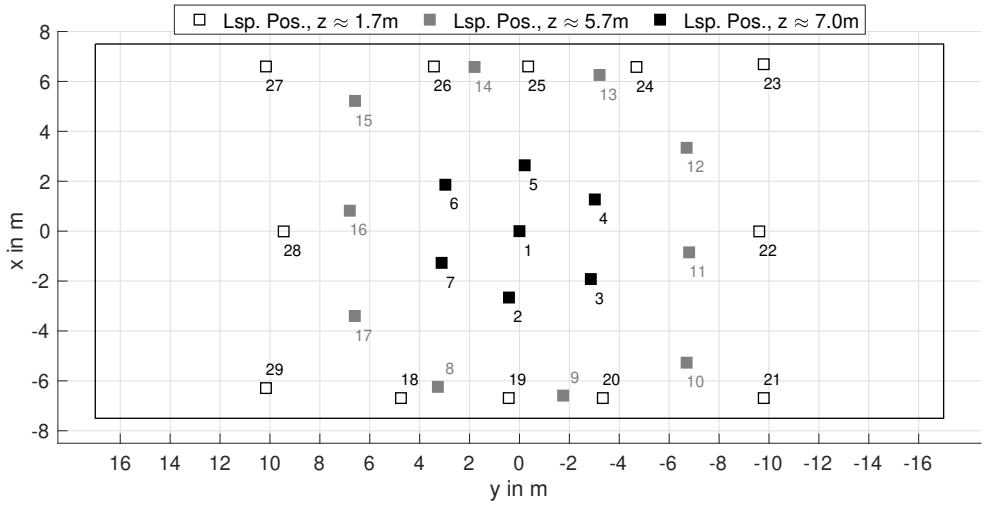
Figure 4.21.: EV_m evaluation in dB at different microphone positions.

Mic. Pos.	$EV_m(\varphi_{\min})$ in dB	$\varphi_{\min}$ in °	$\overline{EV}_m$ in dB	σ in dB
17	-5.5	[-172,-171]	-2.29	1.09
13	-6.2	[-161,-158]	-2.54	1.24
9	-11.8	[-179,-176]	-3.75	2.72
5	≤ -20	[-164,-154],[102,135]	-10.71	6.09

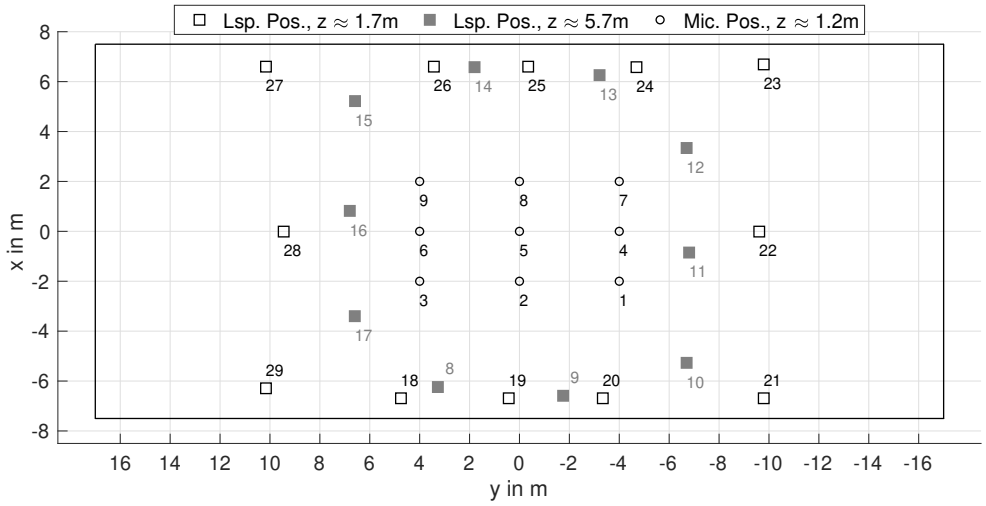
Table 4.4.: The level of the quietest spherical wedges, their directions, averages $\overline{EV}_m$, and standard deviations σ for mic. pos. 17, 13, 9, and 5.

4.3. Exemplary Results for a Concert Hall: György-Ligeti-Hall

The György-Ligeti-Hall is located in the House of Music and Music Theatre (MUMUTH) of the University of Music and Performing Arts Graz. It is equipped with movable loudspeakers of the model CA1001 by Kling&Freitag[3], which are mounted on motorized arms. The arrangement of the 29 loudspeakers can be changed by programming the coordinates of the desired positions into a control interface. With roughly 511 m^2 it is the biggest venue presented in this work.
The used data set consists of 261 FO-ARIRs in total, describing the transmission paths from each of the 29 loudspeakers to the 9 microphone positions[4]. Recordings were performed with two Soundfield ST250 MKII microphones. Fig. 4.22 depicts the loudspeaker and microphone positions.
In order to demonstrate the applicability of the selected measures, the loudspeaker directions are assessed and basic results for the LF_m and the DRR_m are presented.



(a) Loudspeaker positions.



(b) Loudspeaker positions of the two lower rings and microphone positions.

Figure 4.22.: Loudspeaker positions and microphone positions in the György-Ligeti-Hall.

[3]https://www.kling-freitag.de/proclassics/ca-serie/ca-1001/

[4]The measurements were conducted at the 13th and 14th October 2017 by F. Zotter, M. Frank, G. K. Sharma and M. Zaunschirm. The metadata was taken from the documentation written by F. Zotter.

4.3.1. Localization Accuracy

The localization performance in the György-Ligeti-Hall is acceptable. All median errors are under the established thresholds (see sec. 2.1). The biggest mismatches happened for the loudspeakers 6 and 7 in the upper ring ($e_{\text{total},6} = 16.94$, $e_{\text{total},7} = 16.20$). Due to the projection distortion, the mapping error for loudspeaker 1 depicted in fig. 4.23 seems larger than it actually is ($\tilde{e}_{\text{total},1} = 13.97$).

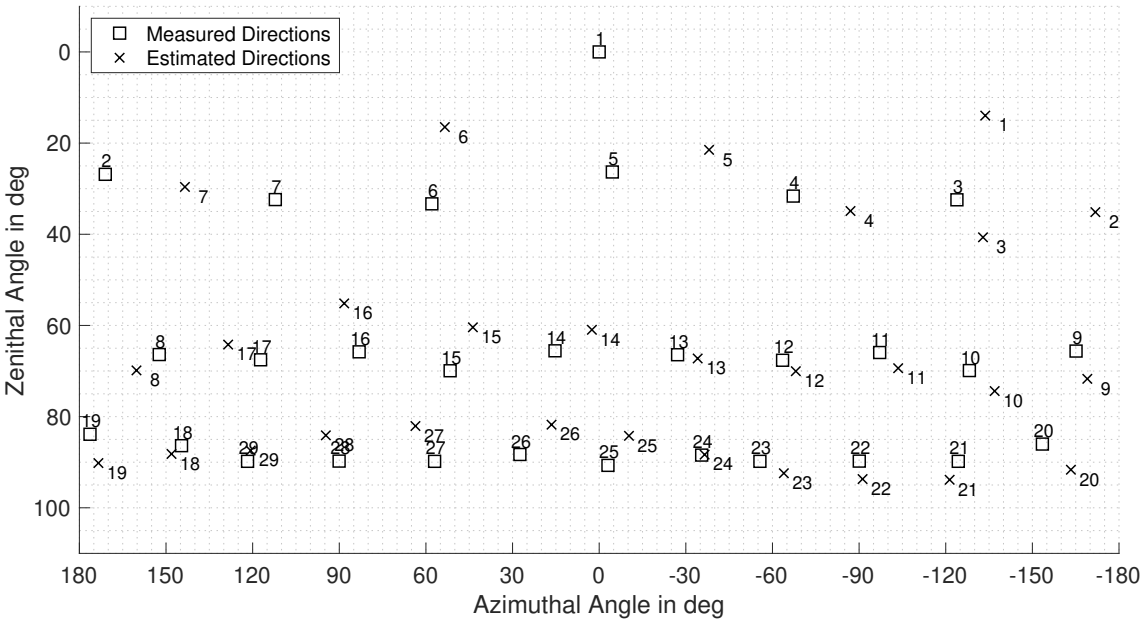


Figure 4.23.: Estimated and measured LSP positions in the György-Ligeti-Hall based on the measurements at mic. pos. 5.

	in °
Median Error $\tilde{e}_{\text{total}}$	9.73 °
Median Azim. Error $\tilde{e}_{\varphi}$ of LSPs 18-29	4.03 °
Median Zen. Error $\tilde{e}_{\vartheta}$	4.61 °

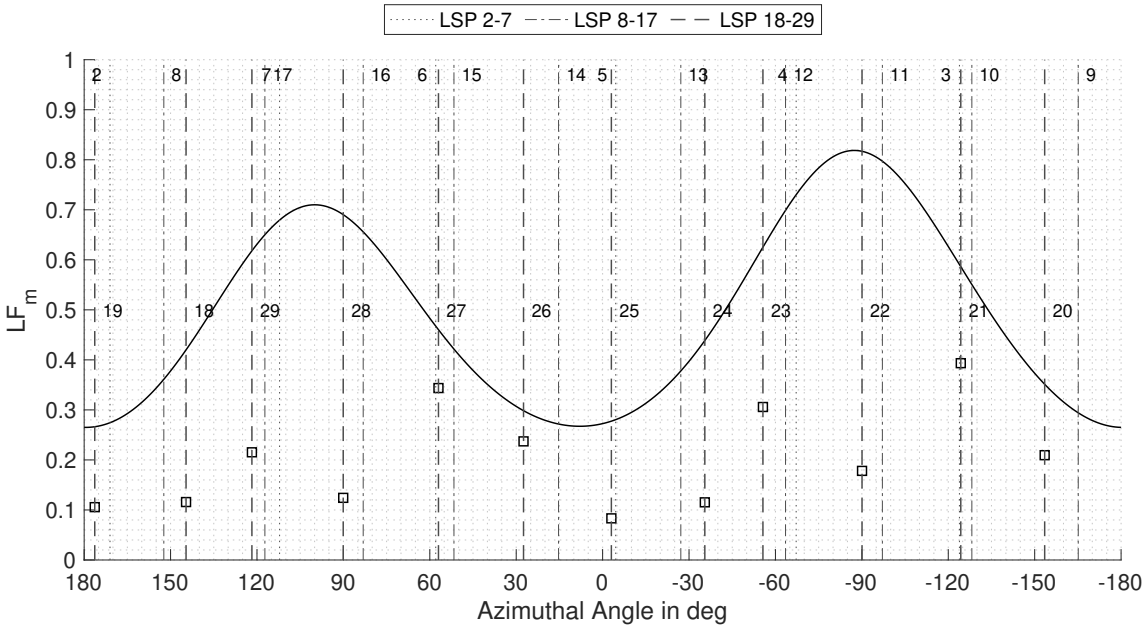
Table 4.5.: Results of the DOA estimation from the off-center mic. pos. 5.

4.3.2. Adapted Lateral Energy Fraction (LF_{m})

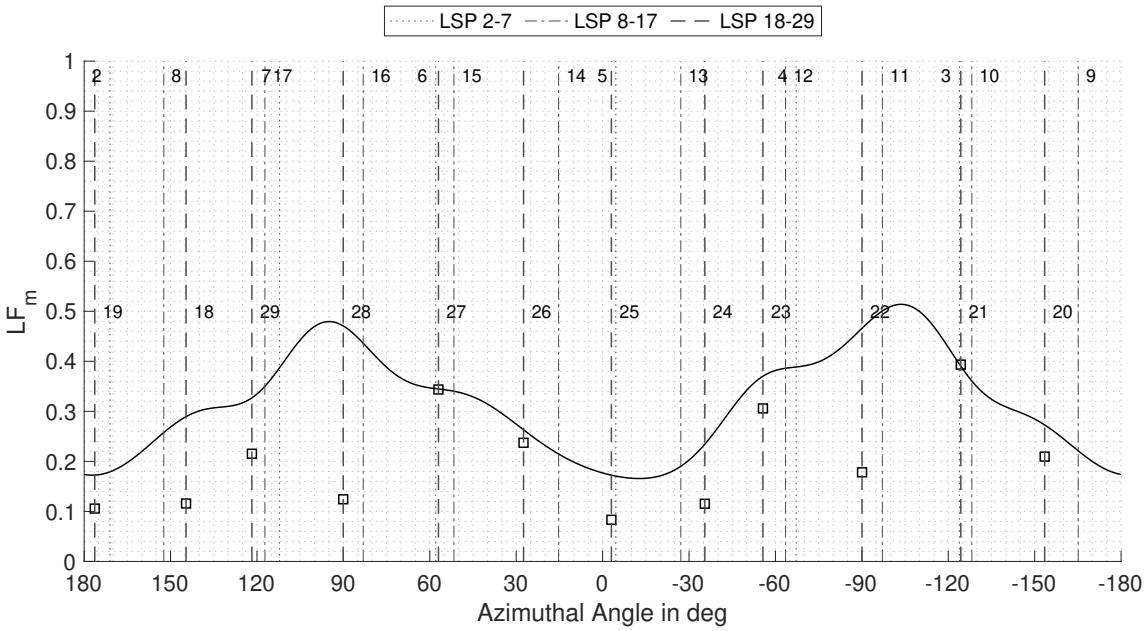
Again, a phantom source is synthesized with different orders of basic-weighted Ambisonics and VBAP at multiple positions on a full circle around the horizon and evaluated in steps of 1 °. How do the major differences in dimension of the room change the perception of the apparent source width?

The assessment of the lateral energy fraction in the György-Ligeti-Hall shows the influence of the room layout in the lower order Ambisonics: The numerous loudspeakers, which participate in 1st-order Ambisonics at the sound reproduction, excite the sound field in the concert hall from several directions and yield a variety of early reflections. When facing the left and right side of the room, the reflections from the front and rear wall are more intense due to the lower distance. This leads to two maximum ASWs around ±90 ° and two minimums near 0 ° and ±180 °.

Fig. 4.24b illustrates that the use of 3rd-order Ambisonics weakens the described fluctuation, since less loudspeaker are active, but the extrema are still clearly recognizable.
The evaluation of the 7th-order Ambisonic scenario (see fig. 4.25a) resembles the results of VBAP (see fig. 4.25b), as seen before during the evaluation of the multi-loudspeaker system in the IEM CUBE. For the notch at loudspeaker 27, where the evaluated LF_m undershoots the value of a single discrete source, no plausible explanation could be found. Loudspeaker 1 is not represented in the following figures, since its position lies at the zenith.

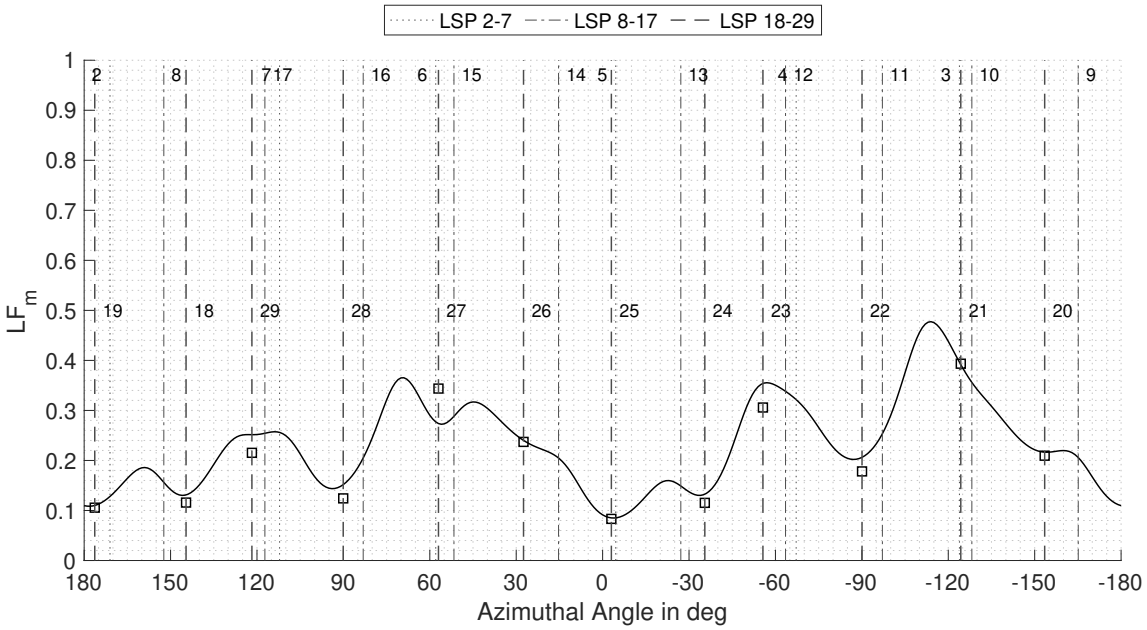


(a) 1st-order Ambisonics.

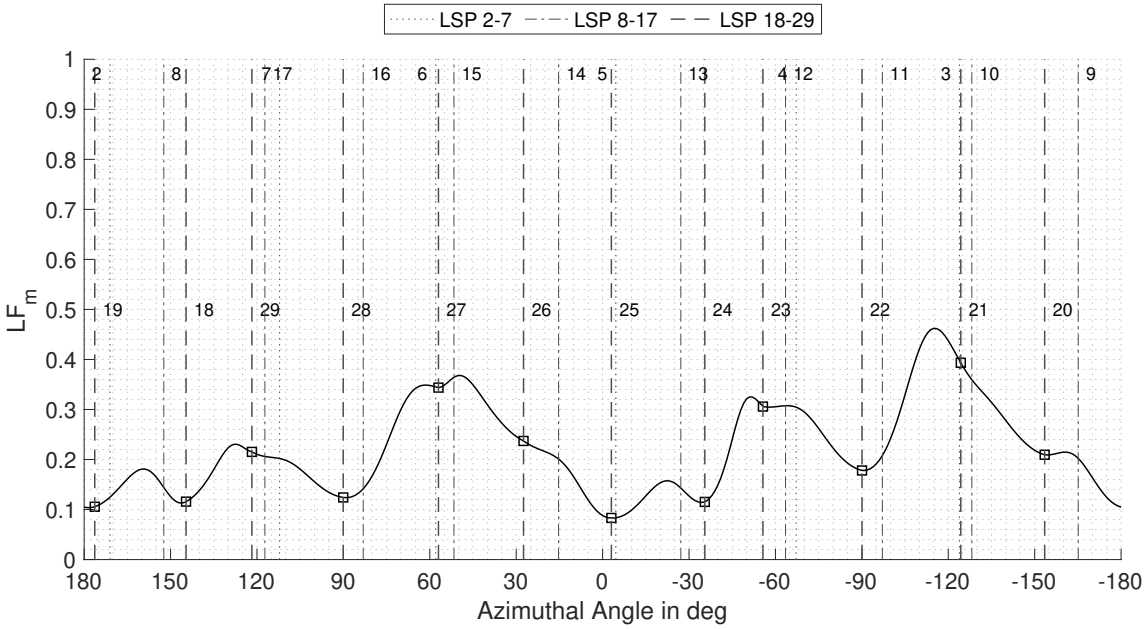


(b) 3rd-order Ambisonics.

Figure 4.24.: LF_m for a phantom source on a circle in the horizontal plane using basic-weighted AllRAD and .



(a) 7th-order Ambisonic.

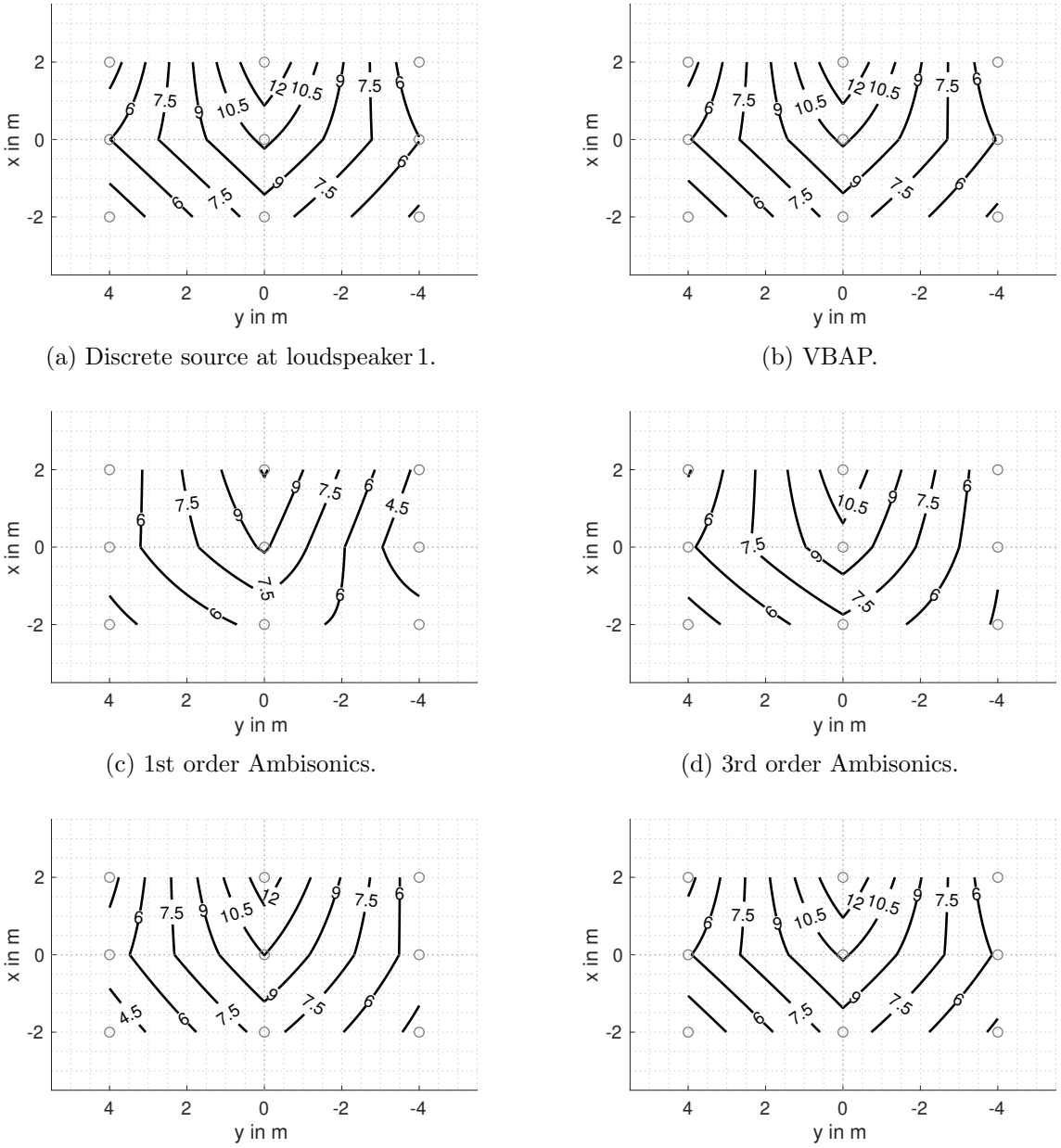


(b) VBAP.

Figure 4.25.: LF_m for a phantom source on a circle in the horizontal plane using basic-weighted AllRAD and VBAP.

4.3.3. Direct-to-Reverberant Energy Ratio (DRR_m)

The DRR_m can also be applied to larger setups in concert halls. Since only the inner square of the listening area is covered by microphone positions, the value of the evaluation is limited. General effects, such as the sharpening of the source width with increasing order and the similarities of 7th-order Ambisonics, LSP 1 and VBAP are still observable. Although the distance to the loudspeaker is larger, the dB-values in the center are clearly higher, compared to the scenario in the IEM CUBE. The narrow angle of radiation of the CA1001 (85 ° horiz. × 55 ° vert.) in contrast to the wider dispersion of the Tannoy System 1200 (90 ° conical) loudspeakers becomes visible. For the assessment of boundary effects and limits of the sweet area, the impulse response measurements at additional microphone positions have to be conducted for further findings.



(a) Discrete source at loudspeaker 1.

(b) VBAP.

(c) 1st order Ambisonics.

(d) 3rd order Ambisonics.

(e) 5th-order Ambisonics.

(f) 7th-order Ambisonics.

Figure 4.26.: DRR_m evaluation for source position at $\boldsymbol{\theta} = [0°, 90°]$ synthesized with basic-weighted AllRAD and VBAP.

4.4. Exemplary Results for a Small Studio: Production Studio

The Production Studio of the IEM has a surface area of approximately 41 m^2. The facility is primarily used by sound engineers and computer musicians for individual work, including mixing and mastering surround sound content. Thus, the loudspeaker setup is designed for a small optimal sweet spot in the middle of the room. The impulse responses to every loudspeaker are measured at this position ($\boldsymbol{\theta}_\mathrm{p} = [0, 0, 1.2]$).
A unique characteristic of the surround sound setup compared to the other two systems is the equal distance of 2.5 m of all loudspeakers to the center. The lower ring of loudspeakers is equipped with Neumann KH310[5] loudspeakers, while the upper hemisphere consists of the older, but identically constructed K+H O300 models.
The room acoustics are improved by absorbers in the corners and mounted at the wall in order to yield a preferably low-reflection environment without dominant room modes and a well-balanced frequency response.

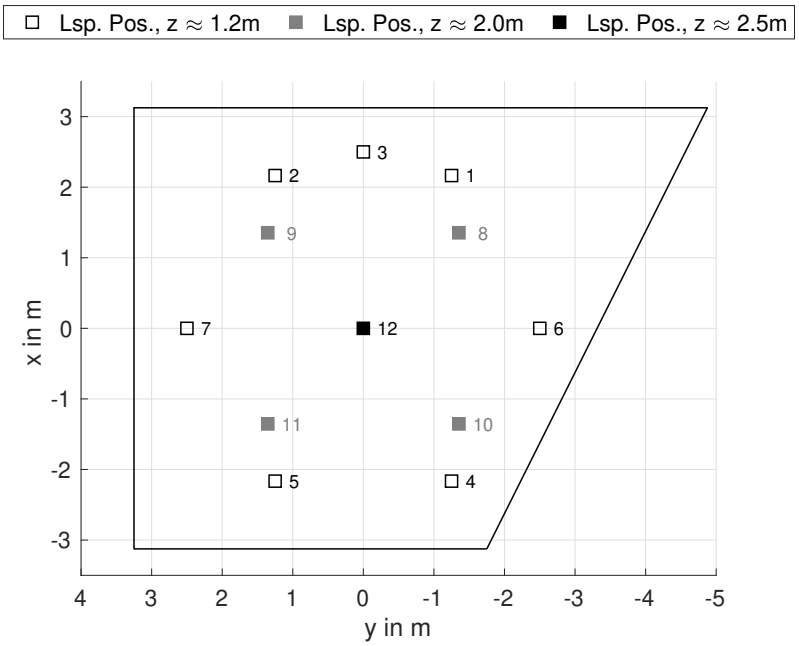


Figure 4.27.: Loudspeaker positions in the Production Studio.

[5]https://de-de.neumann.com/kh-310-a

4.4.1. Localization Accuracy

The loudspeaker localization works very well in the dry sound environment of the acoustically treated studio facility. Comparing the mapping error of the Voice-of-God speaker in the György-Ligeti-Hall (LSP 1) and in the Production Studio (LSP 12), the same direction and magnitude of the bias are noticeable. This could indicate that the mismatch is rather caused by the characteristics of the microphone design, than by the varying predominant acoustics of the two spaces.
The maximum total error e_{total} can be found at loudspeaker 8 and equals approx. 12.20 °.

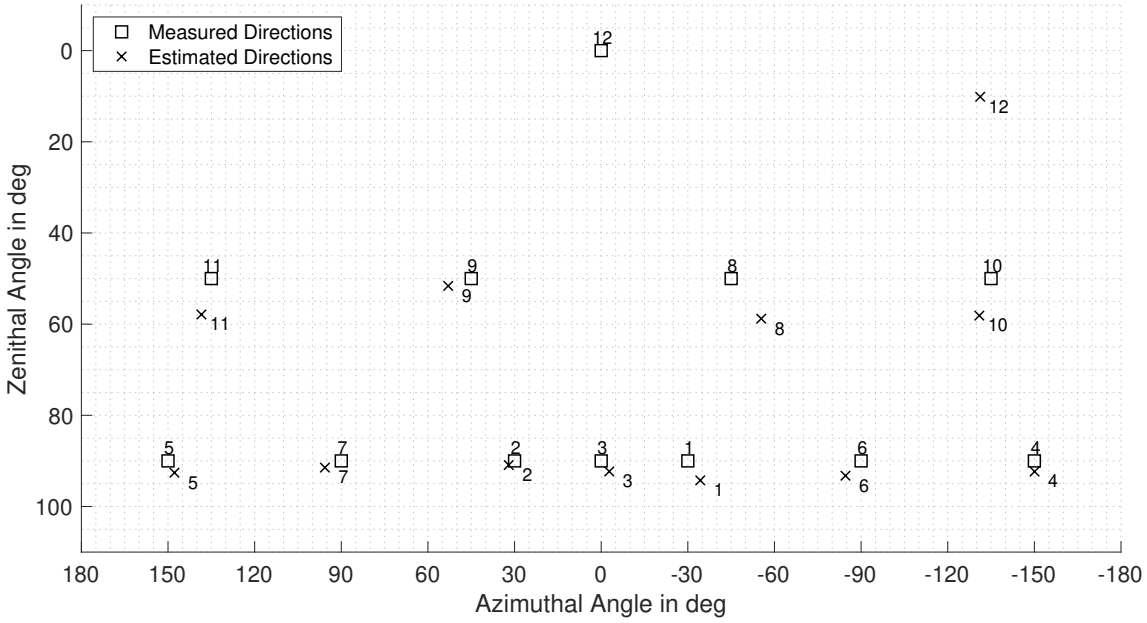


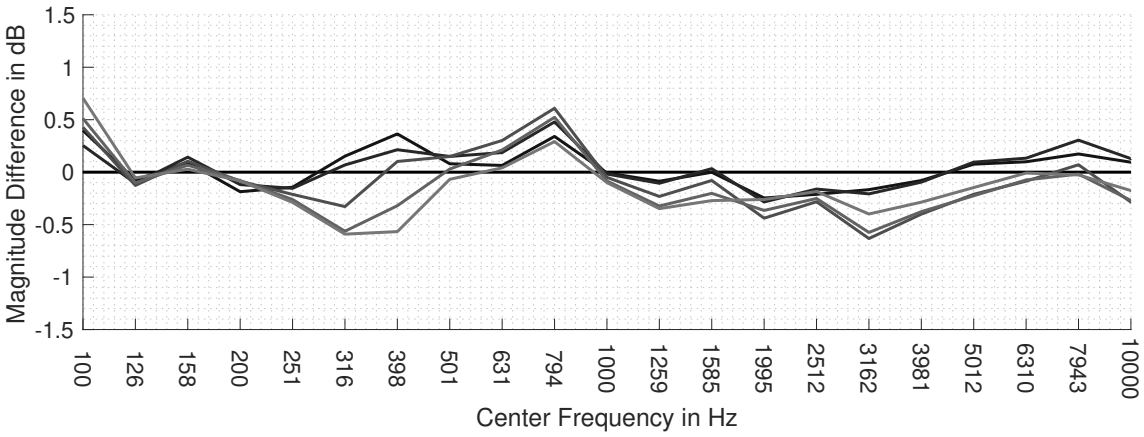
Figure 4.28.: Loudspeaker positions in the Production Studio.

	in °
Median Error $\tilde{e}_{\text{total}}$	6.18 °
Median Azim. Error $\tilde{e}_{\varphi}$ of LSPs 1-7	2.79 °
Median Zen. Error $\tilde{e}_{\vartheta}$	2.92 °

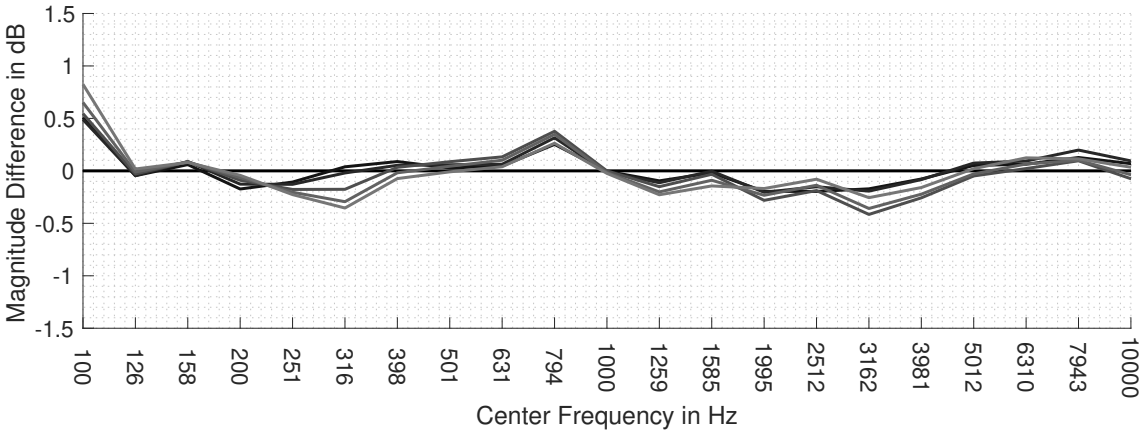
Table 4.6.: Results of the DOA estimation in the Production Studio.

4.4.2. Differences in Coloration (C_m)

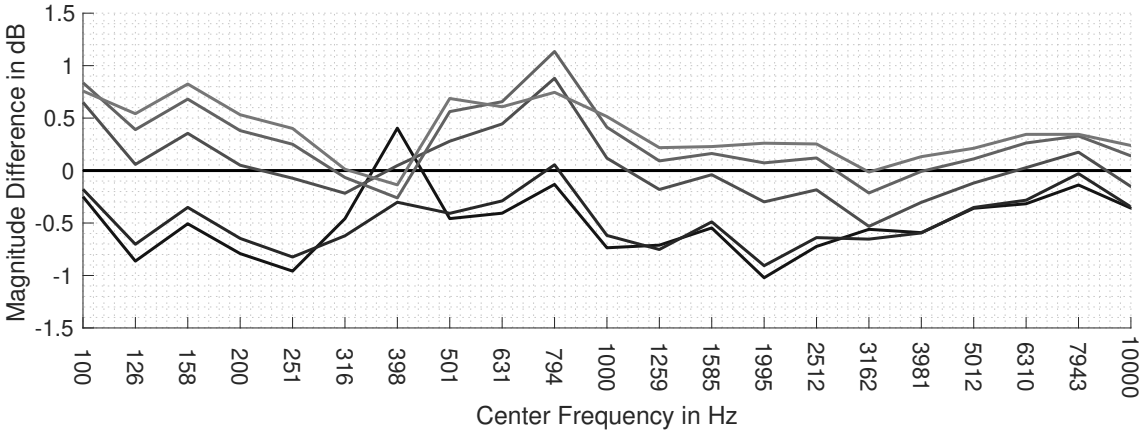
Especially in connection with mixing and mastering tasks, a stable spectrum of moving phantom sources is desired. Phantom source positions between loudspeaker 3 ($\boldsymbol{\theta} = [0, 90°]$) and loudspeaker 2 ($\boldsymbol{\theta} = [30°, 90°]$) are decoded in steps of 5 °.
Overall, the comb filters are less strong than in the IEM CUBE. Again, max-$\boldsymbol{r}_E$ weighting reduces spectral differences. Only VBAP yields potentially perceivable spectral degradations.



(a) AllRAD and basic weighting.



(b) AllRAD and max-$\boldsymbol{r}_E$ weighting.



(c) VBAP.

Figure 4.29.: Spectral difference plots for third-octave bands 7th-order Ambisoncis and VBAP.

4.4.3. Envelopment (EV_m)

Compared to the other loudspeaker arrangements, the hemispherical coverage of the five elevated loudspeakers in the Production Studio is rather sparse. The assessment of the energy level distribution with the EV_m measure shows that the minimum of 7.5 dB is located at the rear between [-165°,165°], where the gap between loudspeaker 4 and 5 is not filled with a loudspeaker of the upper ring. In the frontal section ([-30 °, 30 °]), where the density of loudspeakers is more similar to the one of the horizontal ring in the IEM CUBE, the DRR_m-values improve as well. The standard deviation at the sweet spot equals $\sigma = 1.9$ dB and is worse than the two best of the assessed positions in the IEM CUBE. The mean of the EV_m is -3.6 dB.

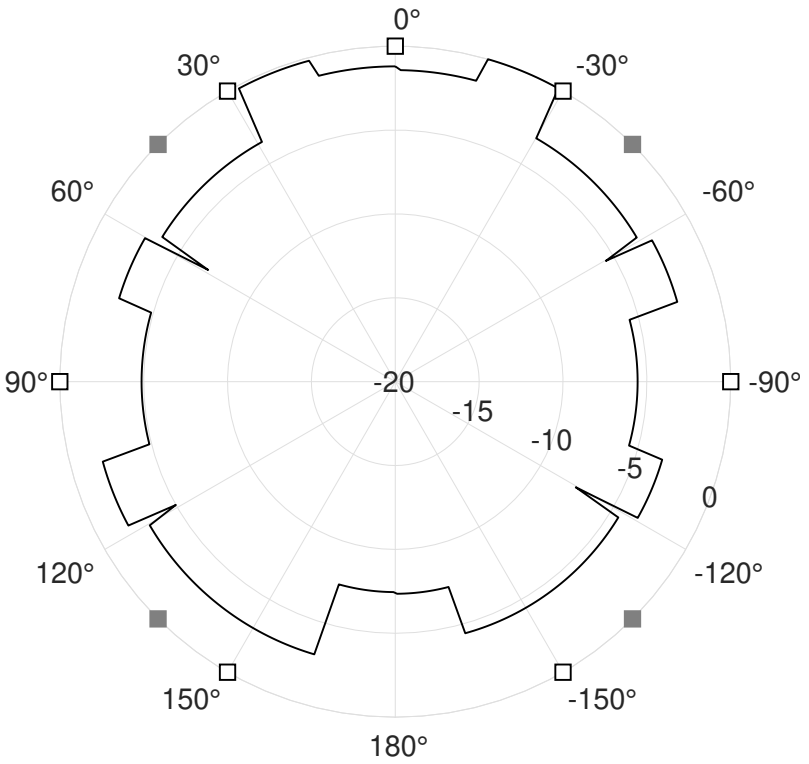


Figure 4.30.: EV_m evaluation in dB at the central microphone position.

Mic. Pos.	$EV_m(\varphi_{\min})$ in dB	$\varphi_{\min}$ in °	$\overline{EV}_m$ in dB	σ in dB
1	-7.5	[-165,165]	-3.6	1.9

Table 4.7.: The level of the quietest spherical wedges, their directions, average $\overline{EV}_m$ and standard deviation σ for the single mic. pos. 1.

5. Conclusions and Outlook

In the first section of this final chapter, the content of this work is summarized and the most important results and conclusions are pointed out.
Open questions and topics for further investigations are examined in the last section.

5.1. A Reasonable Compromise between Simulations and Measurements

This thesis presented a new approach for the evaluation of indoor surround sound setups of hemispherical loudspeaker arrangements based on 1st-order Ambisonic room impulse responses. All findings in this work were gathered by the analysis of microphone signals computed in a virtual work environment. Evaluating the reproduction of phantom sources was the underlying notion for the choice of investigated measures. In total, four measures were discussed:
The direct-to-reverberant energy ratio DRR_m gives information about the distance perception of a phantom source and the coverage of direct sound energy at a listening position can be investigated. An alteration of the window dimensions and a cross-fade enabled the application to multi-loudspeaker systems. Comparing different scenarios and phantom source positions, plausible results could be observed.
The standard lateral energy fraction LF, which estimates the apparent source width, was adapted to the integration times proposed by Frank and the possibility for a rotation of the virtual M/S-setup in the implementation was introduced. The results of the evaluations proved to be in good accordance with the findings in literature. Furthermore, an error estimation for a potential, rotational error yielded that errors up to $10\,^\circ$ caused a relative error of less than 1 dB for the LF_m values.
In order to quantify the spectral differences of moving virtual sources, the coloration measure C of Frank was implemented for the use in the virtual working environment. The results of his previous listening experiments under test conditions could be verified and were supported by the assessment of the practical measurements. AllRAD with max-$\boldsymbol{r}_\mathrm{E}$ weighting achieved great improvements in attenuating comb filters in the IEM CUBE, as well as in the Production Studio.
The envelopment of listeners by spatial reproduction setups is a relatively young topic of research. The developed measure EV_m evaluates to which extend the placement of the loudspeakers ensures a direct sound coverage, independently from the used signal processing method. The average and the standard deviation are potential indicators for the perceived envelopment at a listening position.
The applicability of the measures for differently large and equipped systems was proved by a thorough evaluation of the IEM CUBE and an assessment of selected measures in the György-Ligeti-Hall and the Production Studio.
During the evaluation, advantages and characteristics of the two spatialization methods AllRAD and VBAP were elaborated. While VBAP is captivating by its simplicity and applicability, a lot of problems can be solved by the flexibility of AllRAD: Is the stability of the apparent source width more important than its spread, a low Ambisonic order fulfills this requirement. Is it the other way round, a higher Ambisonic order decreases the ASW

as small as possible. Through the right choice of Ambisonic decoding weights, the comb filters of phantom sources can be reduced further.
A great amount of time was dedicated by the author to developing a robust direction of arrival estimation for loudspeakers, which should replace the measurement of loudspeaker positions by hand. Next to the recommendation of using the central microphone position for the loudspeaker mapping, a methodological least-squares approach for a position estimation procedure was proposed. It promised to yield accurate results but has to be investigated more in depth. The evaluation of the single-microphone-array method in the three differently sized rooms showed overall acceptable estimation results and the detected median errors were comparable to the human localization accuracy. Further improvements could be achieved by measuring the relative direction of the microphone array to a single loudspeaker and aligning the estimations to it.
In summary, it can be stated that the proposed evaluation procedure represents a reasonable compromise between simulations and extensive measurements. Different spatialization methods can now be tested and evaluated regarding different source characteristics, by using an already existing, universal set of 1st-order Ambisonic room impulse measurements. Quality measures, which would require listening experiments, for example the perceived coloration or the lateral energy fraction, can now be assessed solely based on measurements.

5.2. Perspectives in Future Research

With more time on hand, an improvement of the position estimation of loudspeakers can be assumed: Through a clever positioning of the right number of microphone arrays, the least-squares-approach should outperform the single-microphone-method regarding its robustness.
Whereas the LF_m and C_m are already supported by listening experiments, the translation of DRR_m and EV_m values into statements about the subjective perception still has to be conducted. Future listening tests could help to validate whether the presented framework for the DRR_m estimation, with the right choice of time constant for the direct sound interval and the cross-fade, is able to describe the subjective perception of distance and presence.
Furthermore, several interesting questions regarding the envelopment measure EV_m have to be answered: To what extend do the fluctuations, notches, and low-level averages degrade the impression of being enveloped? Are the standard deviation and the average suitable benchmarks for the envelopment perception? Can thresholds for the perception of these characteristics be defined?
In the work, the focus of the sound reproduction methods lied on AllRAD at different Ambisonic orders and weights and VBAP. Obviously, the virtual work environment can easily be adapted to every reproduction method that is also applicable on-site on the physical surround sound setup. Further extensions are conceivable: For example, by applying continuous signals in the input stage and rendering the simulated microphone signals to binaural signals, the sound of the spatialization method in a specific space with its loudspeaker setup can be auralized on headphones for every measured position.
During the thesis, potential shortcomings of the projection accuracy of the used microphone array were suspected. A systematic evaluation could bring more knowledge about the origin of the downward drifts and the mapping errors of auditory events at the zenith.

List of Figures

0.1. Spherical coordinate system used in this thesis and unit sphere, from [3, p.74] 3

1.1. Stereophonic sound system with loudspeaker 1 and 2, aperture angle α and position angle β. 5
1.2. Normalized gains g_1 (dashed line) und g_2 (solid line) of a two-channel sound system (aperture angle $\alpha = 45°$) in dependence of phantom source direction β. 6
1.3. Exemplary Setup for 3D VBAP Using Five Loudspeakers $\boldsymbol{l}_1$-$\boldsymbol{l}_5$, from [7]. . . 8
1.4. Triangle grouping of loudspeakers (LSPs) in the György-Ligeti-Hall with loudspeaker 25 marking the front. 8
1.5. Spherical harmonics of the Ambisonic orders 0 to 5 with positive (bright surface) and negative sign (dark surface), from [16]. 10
1.6. Directivity Patterns for different weightings at Ambisonic orders 1, 3 and 5. 13
1.7. Distribution of 5200 virtual loudspeakers according to the 5200-Point 100-design. 14
1.8. Convex hull and loudspeakers triangles of the system in the IEM CUBE. . . 14

2.1. Length of normalized pseudo-intensity vector $\boldsymbol{I}(t)$ of impulse response of LSP 1 to Mic. Pos. 1 in the IEM CUBE and lower and upper bound of the 1 ms long integration interval. 17

3.1. Signal flow diagram of the virtual working environment utilized in this work. 22
3.2. W-channel the microphone signal and envelopes of the weighting functions $w_{\mathrm{dir}}(t)$ (black) and $w_{\mathrm{rev}}(t)$ (grey). 24
3.3. Standard and adapted evaluation in dB of a single discrete sound source at LSP 1 ($\boldsymbol{\theta}_{\mathrm{p}} = [4.34, 0.12, 1.34]$). 26
3.4. Standard and adapted evaluation in dB of a phantom source encoded in 7th-order Ambisonics between LSP 1 and 2. 26
3.5. Integration windows and individual loudspeaker signals marked with their numbers computed for mic. pos. 20. 27
3.6. Schematic mid-side recording setup on the left side and the corresponding arrangement of two microphones on the right, from [3, p. 5]. 28
3.7. Standard and adapted evaluation of different LSP combinations on the surround sound setup in the IEM CUBE. 30
3.8. Error $e_{LF_m}(\varphi_\Delta)$ in dB for LSP 1 at mic. pos. 17. 31
3.9. Spherical wedge with aperture angle α. 34
3.10. Evaluation of an exemplary playback setup at a microphone position in the center. 34

4.1. Sound energy of the direct part and the first 100 ms of the uncorrected and corrected loudspeaker impulse responses in the IEM CUBE. 38
4.2. Microphone and LSP positions of the measured set of ARIRs (see [50]). . . 39
4.3. Measured and estimated LSP directions from the central mic. pos. 17. . . . 41
4.4. Measured and bias-corrected, estimated LSP directions from the central mic. pos. 17. 41

4.5. Measured and estimated LSP directions from mic. pos. 5. 43
4.6. Course of the normalized length of the PIV $\boldsymbol{I}(t)$ of the impulse response from LSP 12 measured at mic. pos. 5. 44
4.7. Prediction of the perceived azimuth angle φ of the moving phantom source. 45
4.8. Prediction of the perceived zenith angle ϑ of the moving phantom source. . 45
4.9. Prediction of the perceived zenith angle ϑ of the moving phantom source. . 46
4.10. Prediction of the perceived azimuth angle φ of the moving phantom source. 47
4.11. DRR_m contour plot in dB of a discrete sound source at LSP 1. 48
4.12. DRR_m contour plots in dB of a phantom source panned with VBAP. 49
4.13. DRR_m contour plots in dB of phantom sources decoded with basic-weighted Ambisonics. 50
4.14. Normalized loudspeaker gains for all 24 LSPs using different spatialization methods. 51
4.15. LF_m for a phantom source in the horizontal plane using VBAP. 52
4.16. LF_m for a phantom source on a circle in the horizontal plane using AllRAD and basic weighting. 53
4.17. LF_m for a phantom source on a circle in the horizontal plane using AllRAD and basic weighting. 54
4.18. Spectral difference plots in third-octave bands using VBAP. 55
4.19. Spectral difference plots for third-octave bands using 7th-order. 56
4.20. Normalized loudspeaker gains using different spatialization methods and weightings. 57
4.21. EV_m evaluation in dB at different microphone positions. 58
4.22. Loudspeaker positions and microphone positions in the György-Ligeti-Hall. 59
4.23. Estimated and measured LSP positions in the György-Ligeti-Hall based on the measurements at mic. pos. 5. 60
4.24. LF_m for a phantom source on a circle in the horizontal plane using basic-weighted AllRAD and 61
4.25. LF_m for a phantom source on a circle in the horizontal plane using basic-weighted AllRAD and VBAP. 62
4.26. DRR_m evaluation for source position at $\boldsymbol{\theta} = [0°, 90°]$ synthesized with basic-weighted AllRAD and VBAP. 63
4.27. Loudspeaker positions in the Production Studio. 64
4.28. Loudspeaker positions in the Production Studio. 65
4.29. Spectral difference plots for third-octave bands 7th-order Ambisoncis and VBAP. 66
4.30. EV_m evaluation in dB at the central microphone position. 67

A.1. Schematic representation of the localization scenario with microphone arrays at four measurement positions and a single loudspeaker. 88
A.2. e_{LF_m} for LSP 1 and localization matches at mic. pos. 1 to 30. 91
A.3. e_{LF_m} error for all loudspeakers from the central mic. pos. 17. 91
A.4. Sound energy in the direct part and 100 ms-part of the loudspeaker impulse responses in the György-Ligeti-Hall. 92
A.5. Sound energy of the direct part and the first 100 ms of the loudspeaker impulse responses in the Production Studio. 93
A.6. Spectral difference plots for third-octave bands using basic-weighted AllRAD. 94

List of Tables

3.1. Aperture angle of the three investigated pairs of loudspeakers. 30

4.1. Results of the DOA estimation from the center mic. pos. 17. 42
4.2. Results of the DOA estimation from the off-center mic. pos. 5. 42
4.3. Overview over the biased and corrected median of azimuthal and zenithal errors $\tilde{e}_\varphi$ and $\tilde{e}_\vartheta$. 47
4.4. The level of the quietest spherical wedges, their directions, averages $\overline{EV}_m$, and standard deviations σ for mic. pos. 17, 13, 9, and 5. 58
4.5. Results of the DOA estimation from the off-center mic. pos. 5. 60
4.6. Results of the DOA estimation in the Production Studio. 65
4.7. The level of the quietest spherical wedges, their directions, average $\overline{EV}_m$ and standard deviation σ for the single mic. pos. 1. 67

A.1. Ambisonic formats with their channel ordering systems and normalizations. 87

List of Symbols

Scalars

φ	Azimuth Angle
ϑ	Zenith Angle
σ	Standard Deviation
c	Speed of Sound, $343\frac{\mathrm{m}}{\mathrm{s}}$ at $20\,^{\circ}\mathrm{C}$
$\bar{e}$	Mean Error
$\tilde{e}$	Median Error
f_s	Sampling Frequency/Rate
g_i	Gain of Loudspeaker i
w_r, w_τ	Weights of the Extended Energy-Vector
J	Number of Virtual Loudspeakers
E	Sound Energy
L	Number of Loudspeakers
M, m	Ambisonic Degree
N, n	Ambisonic Order

Vectors

$\boldsymbol{\theta}(\varphi, \vartheta)$	Polar Direction Vector
$\hat{\boldsymbol{\theta}}(\varphi, \vartheta)$	Polar Direction Vector of a Virtual Loudspeaker
$\boldsymbol{\theta}_{\mathrm{p}}(x, y, z)$	Cartesian Position Vector
$\boldsymbol{l}_i(x, y, z)$	Cartesian Vector of Loudspeaker i
$\boldsymbol{a}$	Weighting Vector
$\boldsymbol{g}$	Gain Vector
$\boldsymbol{r}_{\mathrm{E}}$	Extended Energy-Vector
$\boldsymbol{I}(t)$	Cartesian Pseudo-Intensity Vector

Matrices

$\boldsymbol{\chi}_N$	Ambisonic Signal of Order N
$\boldsymbol{s}_L$	Signals of L Loudspeakers
$\boldsymbol{D}$	Decoding Matrix
$\hat{\boldsymbol{G}}$	VBAP Rendering Matrix
$\boldsymbol{I}_N$	Identity Matrix of Dimension N
$\boldsymbol{L}$	Geometric Straight Line
$\boldsymbol{M}$	Projection Matrix
$\boldsymbol{P}$	Geometric Plane
$\boldsymbol{R}(m, \varphi)$	Rotation Matrix
$\boldsymbol{Y}_N$	Weighting Matrix
$\hat{\boldsymbol{Y}}_N$	Weighting Matrix for Virtual Loudspeakers

Functions

δ_{m0}	Kronecker Delta
$\delta(t)$	Dirac Delta function
$s(t)$	Audio Signal

$w_{\mathrm{dir}}(t), w_{\mathrm{rev}}(t)$	Weighting-Envelope Function
$\Phi_m(\varphi)$	Azimuthal Part of Y_n^m
$\Theta_n^m(\vartheta)$	Zenithal Part of Y_n^m
$N_n^{\lvert m\rvert}$	Normalization term of $\Theta_n^m(\vartheta)$
$P_n^{\lvert m\rvert}$	Legendre Polynom of Order n and Degree m
$W(t), X(t), Y(t), Z(t)$	Channels of the First Ambisonic Order
Y_n^m	Spherical Harmonic of Order n and Degree m

Operators

$\langle \cdot \rangle$	Scalar Product in the Euclidean Space
$\mathrm{diag}\{ \cdot \}$	Diagonal Matrix
$\max\{ \cdot \}$	Maximum
$\min\{ \cdot \}$	Minimum

List of Abbreviations

C_m	Adapted Degree of Coloration
DRR	Direct-to-Reverberation Energy Ratio
DRR_m	Adapted Direct-to-Reverberation Energy Ratio
EV_m	Adapted Envelopment
LEV	Listener Envelopment
LF	Lateral Energy Fraction
LF_m	Adapted Lateral Energy Fraction
3D	Three-Dimensional
ACN	Ambisonic Channel Number
AllRAD	All-round Ambisonic Decoding
ARIR	Ambisonic Room Impulse Response
ASW	Apparent/Auditory Source Width
DAW	Digital Audio Workstation
DOA	Direction of Arrival
EPAD	Energy-Preserving Ambisonic Decoding
FO-ARIR	First-Order Ambisonics Room Impulse Response
FOA	First-Order Ambisonics
HOA	Higher-Order Ambisonics
ICLD	Interchannel Level Differences
ICTD	Interchannel Time Differences
IEM	Institute for Electronic Music and Acoustics
ILD	Interaural Level Differences
ITD	Interaural Time Differences
LSP	Loudspeaker
M/S	Mid-Side (Setup)
MAD	Mode Matching Decoder
MUMUTH	House of Music and Music Theatre
N3D	Full 3D Normalization
PIV	Pseudo-Intensity Vector
SAD	Sampling Decoder
SDK	Software Development Kit
SDM	Spatial Decomposition Method

SN3D	Schmidt Semi-Normalization
SVD	Singular Value Decomposition
VBAP	Vector-Base Amplitude Panning
VR	Virtual Reality
WFS	Wave Field Synthesis

Bibliography

[1] ISO/IEC 23008-3:2019, *Information technology — high efficiency coding and media delivery in heterogeneous environments — part 3: 3d audio*, 2019. [Online]. Available: https://www.iso.org/standard/74430.html.

[2] ETSI TS 103 491, *Dts-uhd audio format; delivery of channels, objects and ambisonic sound fields*, 2019. [Online]. Available: https://www.techstreet.com/standards/etsi-ts-103-491?product_id=2103224.

[3] F. Zotter and M. Frank, *Ambisonics, a practical 3D audio theory for recording, studio production, sound reinforcement, and virtual reality*, 1st ed. Springer Nature, 2019. DOI: 10.1007/978-3-030-17207-7.

[4] K. Wendt, "Das Richtungshören bei der Überlagerung zweier Schallfelder bei Intensitäts- und Laufzeitstereophonie," Aachen, Techn. Hochsch., Diss., 1964, Ph.D. dissertation, Aachen, 1964. [Online]. Available: https://publications.rwth-aachen.de/record/79513.

[5] A. J. Berkhout, "A holographic approach to acoustic control," *Journ. Audio Eng. Soc*, vol. 36, no. 12, pp. 977–995, 1988. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=5117.

[6] A. J. Berkhout, D. de Vries, and P. Vogel, "Acoustic control by wave field synthesis," *Journ. Acoust. Soc. Am.*, vol. 93, no. 5, pp. 2764–2778, 1993. DOI: 10.1121/1.405852.

[7] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *Journ. Audio Eng. Soc*, vol. 45, no. 6, pp. 456–466, 1997. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=7853.

[8] J. Blauert, *Spatial Hearing, The Psychophysics of Human Sound Localization*, 2nd ed. The MIT Press, 1997. DOI: 10.1121/1.392109.

[9] D. M. Leakey, "Some measurement on the effect of interchannel intensity and time differences in two channel systems," *Journ. Acoust. Soc. Am.*, vol. 31, no. 7, pp. 977–986, 1959. DOI: 10.1121/1.1907824.

[10] B. B. Bauer, "Phasor analysis of some stereophonic phenomena," *Journ. Acoust. Soc. Am.*, vol. 33, no. 1, pp. 1536–1540, Nov. 1961. DOI: 10.1121/1.1908492.

[11] F. Zotter and M. Frank, "All-Round Ambisonic Panning and Decoding," *J. Audio Eng. Soc.*, vol. 60, no. 10, pp. 807–820, 2012. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=16554.

[12] C. B. Barber, D. P. Dopkin, and H. Huhdanpaa, "The quickhull algorithm for convex hulls," *ACM Transactions on Mathematical Software*, vol. 22, no. 4, pp. 469–483, Dec. 1996. DOI: 10.1145/235815.235821.

[13] D. Malham, "Ambisonics - a technique for low cost, high precision, three dimensional sound diffusion," *International Computer Music Conference*, 1990. [Online]. Available: http://hdl.handle.net/2027/spo.bbp2372.1990.030.

[14] M. A. Gerzon, "Ambisonics. part two: Studio techniques," *Studio Sound*, vol. 17, 24ff, Aug. 1975.

[15] ——, "The Design of Precisely Coincident Microphone Arrays for Stereo and Surround Sound," in *Audio Engineering Society Convention 50*, Mar. 1975. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=2466.

[16] F. Zotter. (2013). "Spherical Harmonics up to Ambisonic order 5 as commonly displayed, sorted by increasing Ambisonic Channel Number (ACN), aligned for symmetry.," [Online]. Available: https://en.wikipedia.org/wiki/Ambisonic_data_exchange_formats#/media/File:Spherical_Harmonics_deg5.png (visited on 10/05/2021).

[17] M. Chapman, W. Ritsch, T. Musil, I. Zmölnig, H. Pomberger, F. Zotter, and A. Sontacchi, "A Standard for Interchange of Ambisonic Signal Sets: Including a File Standard with Metadata," in *Ambisonics Symposium, Graz*, 2009. [Online]. Available: https://iem.kug.ac.at/fileadmin/media/iem/projects/2009/ambixchange09.pdf.

[18] D. Malham, "Higher order ambisonic systems, abstracted from: Space in music - music in space," Ph.D. dissertation, University of York, 2003.

[19] J. Daniel, "Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia," Ph.D. dissertation, Université Paris 6, 2001.

[20] C. Nachbar, F. Zotter, E. Delelfie, and A. Sontacchi, "AmbiX - A suggested Ambisonics Format," *Ambisonic Symposium, Lexington, KY*, 2011. [Online]. Available: https://iem.kug.ac.at/fileadmin/media/iem/projects/2011/ambisonics11_nachbar_zotter_sontacchi_deleflie.pdf.

[21] M. Frank and F. Zotter, "Spatial impression and directional resolution in the reproduction of reverberation," *Fortschritte der Akustik - DEGA*, 2016.

[22] ——, "Exploring the perceptual sweet area in ambisonics," in *Audio Engineering Society Convention 142*, May 2017. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=18604.

[23] V. Pulkki, "Spatial sound reproduction with directional audio coding," *Journ. Audio Eng. Soc*, vol. 55, no. 6, pp. 503–516, 2007. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=14170.

[24] S. Tervo, J. Pätynen, A. Kuusinen, and T. Lokki, "Spatial decomposition method for room impulse responses," *Journ. Audio Eng. Soc*, vol. 61, no. 1/2, pp. 17–28, 2013. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=16664.

[25] L. McCormack, V. Pulkki, A. Politis, O. Scheuregger, and M. Marschall, "Higher-Order Spatial Impulse Response Rendering: Investigating the Perceived Effects of Spherical Order, Dedicated Diffuse Rendering, and Frequency Resolution," *J. Audio Eng. Soc.*, vol. 68, no. 5, pp. 338–354, 2020. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=20852.

[26] M. Zaunschirm, M. Frank, and F. Zotter, "Binaural Rendering with Measured Room Responses: First-Order Ambisonic Microphone vs. Dummy Head," *Applied Sciences*, vol. 10, no. 5, 2020. DOI: 10.3390/app10051631.

[27] F. Zotter, H. Pomberger, and M. Noisternig, "Energy-Preserving Ambisonic Decoding," *Acta Acustica united with Acustica*, vol. 98, no. 1, pp. 37–47, 2012. DOI: 10.3813/AAA.918490.

[28] M. A. Poletti, "A unified theory of horizontal holographic sound systems," *Journ. Audio Eng. Soc*, vol. 48, no. 12, pp. 1155–1182, 2000. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=12033.

[29] ——, "Three-Dimensional Surround Sound Systems Based on Spherical Harmonics," *Journ. Audio Eng. Soc*, vol. 53, no. 11, pp. 1004–1025, 2005. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=13396.

[30] F. Zotter, H. Pomberger, and M. Frank, "Comparison of energy-preserving and all-round ambisonic decoders," in *AIA-DAGA 2013, International Conference on Acoustics, Merano*, Deutsche Gesellsch. f. Akustik, Jun. 2013, ISBN: 978-3939296058.

[31] M. Frank, F. Zotter, and A. Sontacchi, "Localization experiments using different 2d ambisonics decoders," *25th VDT International Convention*, Nov. 2008.

[32] M. Gräf and D. Potts, "On the computation of spherical designs by a new optimization approach based on fast spherical fourier transforms," *Numerische Mathematik*, no. 119, pp. 699–724, 2011. DOI: 10.1007/s00211-011-0399-7.

[33] Recommendation ITU-R BS.2051-0, *Advanced sound system for programme production*, Geneva, 2014. [Online]. Available: https://www.itu.int/rec/R-REC-BS.2051/en.

[34] M. Romanov, M. Frank, F. Zotter, and F. Nixon, "Manipulations improving amplitude panning on small standard loudspeaker arrangements for surround with height," *29th VDT International Convention*, Nov. 2016.

[35] J. Blauert, "Sound localization in the median plane," *Acta Acustica united with Acustica*, vol. 22, pp. 205–213, Nov. 1969.

[36] M. B. Gardner and R. S. Gardner, "Problem of localization in the median plane: Effect of pinnae cavity occlusion," *The Journal of the Acoustical Society of America*, vol. 53, no. 2, pp. 400–408, 1973. DOI: 10.1121/1.1913336.

[37] V. R. Algazi, C. Avendano, and R. O. Duda, "Elevation localization and head-related transfer function analysis at low frequencies," *Journ. Acoust. Soc. Am.*, vol. 109, no. 3, pp. 1110–1122, 2001. DOI: 10.1121/1.1349185.

[38] V. Pulkki, "Localization of amplitude-panned virtual sources ii: Two- and three-dimensional panning," *Journ. Audio Eng. Soc*, vol. 49, no. 9, pp. 753–767, 2001. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=10179.

[39] F. Wendt, "Beurteilung von Phantom Schallquellen vertikaler Anordnungen," Master Thesis, University of Music and Performing Arts Graz, Graz, 2013. [Online]. Available: https://iem.kug.ac.at/fileadmin/media/iem/projects/2012/wendt.pdf.

[40] F. Wendt, M. Frank, and F. Zotter, "Panning with Height on 2,3, and 4 Loudspeakers," in *Proceedings of ICSA 2014 in Erlangen*, 2014, pp. 184–188, ISBN: 978-3-98 12830-4-4.

[41] A. W. Bronkhorst, "Localization of real and virtual sound sources," *Journ. Acoust. Soc. Am.*, vol. 98, no. 5, pp. 2542–2553, 1995. DOI: 10.1121/1.413219.

[42] P. Zahorik, D. Brungart, and A. Bronkhorst, "Auditory distance perception in humans: A summary of past and present research," *Acta Acustica united with Acustica*, vol. 91, pp. 409–420, May 2005.

[43] J. C. Makous and J. C. Middlebrooks, "Two-dimensional sound localization by human listeners," *The Journal of the Acoustical Society of America*, vol. 87, no. 5, pp. 2188–2200, 1990. DOI: 10.1121/1.399186.

[44] S. Carlile, P. Leong, and S. Hyams, "The nature and distribution of errors in sound localization by human listeners," *Hearing Research*, vol. 114, no. 1, pp. 179–196, 1997. DOI: https://doi.org/10.1016/S0378-5955(97)00161-5.

[45] M. Frank, "Phantom sources using multiple loudspeakers in the horizontal plane," Ph.D. dissertation, University of Music and Performing Arts Graz, 2013.

[46] M. Frank and F. Zotter, "Extension of the generalized tangent law for multiple loudspeakers," *Fortschritte der Akustik - DAGA*, 2017.

[47] P. Stitt, S. Bertet, and M. Van Walstijn, "Extended Energy Vector Prediction of Ambisonically Reproduced Image Direction at Off-Center Listening Positions," *Journ. Audio Eng. Soc.*, vol. 64, no. 5, pp. 299–310, May 2016. [Online]. Available: `https://www.aes.org/e-lib/browse.cfm?elib=18135`.

[48] P. Stitt, S. Bertet, and M. Van Walstijn, "Off-Center Listening with Third-Order Ambisonics: Dependence of Perceived Source Directions on Signal Type," *Journ. Audio Eng. Soc.*, vol. 65, no. 3, pp. 188–197, Mar. 2017. [Online]. Available: `https://www.aes.org/e-lib/browse.cfm?elib=18554`.

[49] E. Kurz and M. Frank, "Prediction on the listening area based on the energy vector," *Proc. 4th Int. Conf. on Spatial Audio (ICSA), Graz*, 2017.

[50] K. Müller and F. Zotter. (2020). "CUBE B-format RIR dataset (Soundfield ST450 MKII)," [Online]. Available: `https://phaidra.kug.ac.at/view/o:104435` (visited on 02/09/2021).

[51] A. W. Bronkhorst and T. Houtgast, "Auditory distance perception in rooms," *Nature*, vol. 397, no. 6719, pp. 517–520, 1999. DOI: `10.1038/17374`.

[52] S. Csadi, F. M. Boland, L. Ferguson, H. O'Dwyer, and E. Bates, "Direct to reverberant ratio measurements in small and mid-sized rooms," *Journ. Audio Eng. Soc.*, Mar. 2019. [Online]. Available: `https://www.aes.org/e-lib/browse.cfm?elib=20407`.

[53] D. Griesinger, "The importance of the direct to reverberant ratio in the perception of distance, localization, clarity, and envelopment," in *Audio Engineering Society Convention 126*, May 2009. [Online]. Available: `http://www.aes.org/e-lib/browse.cfm?elib=14920`.

[54] A. Kuusinen and T. Lokki, "Investigation of auditory distance perception and preferences in concert halls by using virtual acoustics," *Journ. Acoust. Soc. Am.*, vol. 138, no. 5, pp. 3148–3159, 2015. DOI: `10.1121/1.4935388`.

[55] M. Barron and A. H. Marshall, "Spatial impression due to early lateral reflections in concert halls: The dreivation of a physical measure," *Journ. of Sound and Vibration*, vol. 77, no. 2, pp. 211–232, 1981. DOI: `10.1016/S0022-460X(81)80020-X`.

[56] T. Hidaka, L. L. Beranek, and T. Okano, "Interaural cross-correlation, lateral fraction and low and high frequency sound levels as measures of acoustical quality in concert halls," *Journ. Acoust. Soc. Am.*, vol. 98, no. 2, pp. 988–1007, Aug. 1995. DOI: `10.1121/1.414451`.

[57] M. Frank, "Source width of frontal phantom sources: Perception, measurement, and modeling," *Archives of Acoustics*, vol. vol. 38, no. No 3, pp. 311–319, 2013. DOI: `10.2478/aoa-2013-0038`.

[58] T. Cox, W. Davies, and Y. Lam, "The sensitivity of listeners to early sound field changes in auditoria," *Acta Acustica united with Acustica*, vol. 79, pp. 27–41, Jul. 1993.

[59] M. Blau, "Difference limens for measures of apparent source width," *Forum Acusticum, Sevilla, Spain*, 2002.

[60] F. Rumsey, S. Zieliński, R. Kassier, and Bech, "On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio quality," *Journ. Acoust. Soc. Am.*, vol. 118, no. 2, pp. 968–976, Aug. 2005. DOI: 10.1121/1.1945368.

[61] M. Karjalainen, E. Piirilä, A. Järvinen, and J. Hyuopaniemi, "Comparison of Loudspeaker Equalization Methods Based on DSP Techniques," *Journ. Audio Eng. Soc.*, vol. 47, no. 1, pp. 14–31, Jan. 1999. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=12117.

[62] EN61260-1:2014, *Electroacoustics – octave-band and fractional-octave-band filters – part 1: Specifications*, 2015. [Online]. Available: https://shop.austrian-standards.at/action/de/public/details/547588/OEVE_OENORM_EN_61260-1_2015_08_01.

[63] J. Berg, "The contrasting and conflicting definitions of envelopment," *Journ. Audio Eng. Soc.*, May 2009. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=15004.

[64] T. Hidaka, T. Okano, and L. Beranek, "Interaural cross correlation (iacc) as a measure of spaciousness and envelopment in concert halls," *Journ. Acoust. Soc. Am.*, vol. 92, no. 4, p. 2469, Oct. 1992. DOI: 10.1121/1.404472.

[65] J. S. Bradley and G. A. Soulodre, "Acoustics 1995: Listener envelopment: An essential part of good concert hall acoustics," *Journ. Acoust. Soc. Am.*, vol. 99, no. 1, pp. 22–22, 1996. DOI: 10.1121/1.414533.

[66] M. Morimoto, "The role of rear loudspeakers in spatial impression," *Journ. Audio Eng. Soc.*, Sep. 1997. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=7225.

[67] A. Wakuda, H. Furuya, K. Fujimoto, K. Isogai, and K. Anai, "Effects of arrival direction of late sound on listener envelopment," *Acoustical Science and Technology*, vol. 24, no. 4, pp. 179–185, 2003. DOI: 10.1250/ast.24.179.

[68] G. A. Soulodre, M. C. Lavoie, and S. G. Norcross, "Objective measures of listener envelopment in multichannel surround systems," *Journ. Audio Eng. Soc.*, vol. 51, no. 9, pp. 826–840, Sep. 2003. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=12205.

[69] K. Hiyama, S. Komiyama, and K. Hamasaki, "The minimum number of loudspeakers and its arrangement for reproducing the spatial impression of diffuse sound field," in *Audio Engineering Society Convention 113*, Oct. 2002. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=11272.

[70] ISO/IEC 60651:1979, *Sound level meters*, 1979. [Online]. Available: https://webstore.iec.ch/publication/17086.

A. Appendix

A.1. Overview over Ambisonic Formats

order	degree	ambiX			B-format	
n	m	ANC	SN3D	N3D	FuMa	FuMa weights
0	0	0	0.282	0.282	W	$\frac{1}{\sqrt{2}}$
1	-1	1	0.282	0.489	Y	1
1	0	2	0.282	0.489	Z	1
1	-1	3	0.282	0.489	X	1
2	-2	4	0.081	0.182	V	$\frac{2}{\sqrt{3}}$
2	-1	5	0.163	0.364	T	$\frac{2}{\sqrt{3}}$
2	0	6	0.282	0.631	R	1
2	1	7	0.163	0.364	S	$\frac{2}{\sqrt{3}}$
2	2	8	0.081	0.182	U	$\frac{2}{\sqrt{3}}$
3	-3	9	0.015	0.039	Q	$\sqrt{\frac{8}{5}}$
3	-2	10	0.036	0.096	O	$\frac{3}{\sqrt{5}}$
3	-1	11	0.115	0.305	M	$\sqrt{\frac{45}{32}}$
3	0	12	0.282	0.746	K	1
3	1	13	0.115	0.305	L	$\sqrt{\frac{45}{32}}$
3	2	14	0.036	0.096	N	$\frac{3}{\sqrt{5}}$
3	3	15	0.015	0.039	P	$\sqrt{\frac{8}{5}}$
⋮	⋮	⋮	⋮	⋮	-	-

Table A.1.: Ambisonic formats with their channel ordering systems and normalizations.

A.2. Approach of an Improved Position Estimation

Fig. A.1 illustrates the idea behind the proposal for a more robust estimation procedure. Using the direction vectors from the loudspeaker estimation (grey arrows), a set of lines (grey, dashed lines) can be constructed based on the position vector of the microphone array. Under the assumption of non-systematic, but random errors in measurements, the acoustic center of loudspeaker (black point) should coincide with the point of the smallest distance to all lines (grey point).

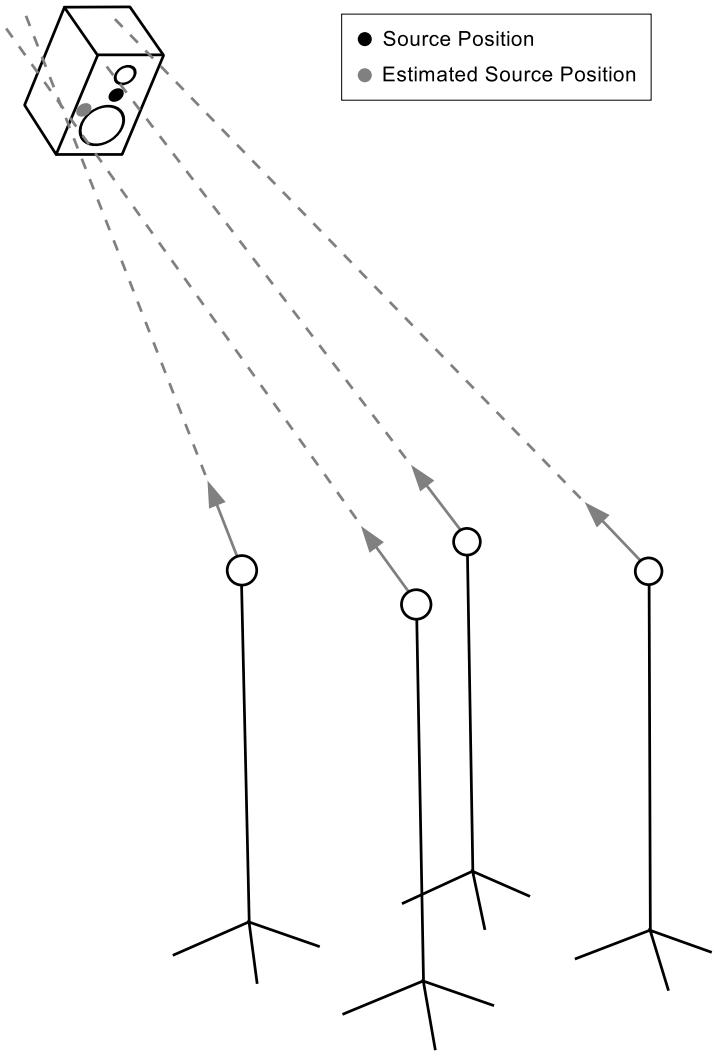


Figure A.1.: Schematic representation of the localization scenario with microphone arrays at four measurement positions and a single loudspeaker.

Method of Least Squares Calculating the point of the smallest distance to a set of M lines is a typical least square problem and can be solved by minimizing the sum of squared distances using the Moore-Penrose inverse $\boldsymbol{A}^{+}$. All of the following equations, vectors, and matrices are formulated in the Cartesian coordinate system.
The well-known position of the microphone array i serves as the position vector $\boldsymbol{m}_i$ and the direction vector $\boldsymbol{\theta}_i$ (calculated with the method presented in sec. 2.2) points to the

estimated direction of the loudspeaker. These two variables define a line $\boldsymbol{L}_i$ in $\mathbb{R}^3$

$$\boldsymbol{L}_i = \boldsymbol{\theta}_i t + \boldsymbol{m}_i \quad . \tag{A.1}$$

In order to calculate the *shortest* distance from an arbitrary point x (with the corresponding position vector $\boldsymbol{x}$) to the line $\boldsymbol{L}_i$, a the point is projected into the plane $\boldsymbol{P}_i$ which includes the origin and is orthogonal to $\boldsymbol{\theta}_i$.

$$\hat{\boldsymbol{x}} = O(\boldsymbol{x}) = (\boldsymbol{I}_N - \boldsymbol{\theta}_i \boldsymbol{\theta}_i^T)\boldsymbol{x}\,. \tag{A.2}$$

For a better readability, this matrix multiplication, referred to as $O(\boldsymbol{x})$ in here, is derived in the subsequent paragraph.
As follows, the smallest distance of the point $\boldsymbol{x}$ and the line $\boldsymbol{L}_i$ is the distance between the projected position vector $\hat{\boldsymbol{x}}$ and the point of intersection $\boldsymbol{s}_i$ of the plane $\boldsymbol{P}_i$ and the line $\boldsymbol{L}_i$. $\boldsymbol{s}_i$ is given by the equation

$$\boldsymbol{s}_i = \boldsymbol{M}\boldsymbol{m}_i = (\boldsymbol{I}_N - \boldsymbol{\theta}_i \boldsymbol{\theta}_i^T)\boldsymbol{m}_i \quad . \tag{A.3}$$

The smallest distance d_i from point $\boldsymbol{x}$ to line $\boldsymbol{L}_i$ can be written as

$$d_i = ||\hat{\boldsymbol{x}} - \boldsymbol{s}_i|| = ||\boldsymbol{M}\boldsymbol{x} - \boldsymbol{s}_i|| = ||(\boldsymbol{I}_N - \boldsymbol{\theta}_i \boldsymbol{\theta}_i^T)\boldsymbol{x} - \boldsymbol{s}_i||\,. \tag{A.4}$$

Therefore, the sum of squared distances D of a point x to the set of i lines equals

$$D = \sum_{i=1}^{M} d_i^2 = \sum_{i=1}^{M} ||(\boldsymbol{I}_N - \boldsymbol{\theta}_i \boldsymbol{\theta}_i^T)\boldsymbol{x} - \boldsymbol{s}_i||^2\,. \tag{A.5}$$

This equation can be formed into the well-known structure

$$||\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}||^2 = \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} + 2\boldsymbol{b}^T\boldsymbol{x} + c \tag{A.6}$$

with

$$\boldsymbol{A} = \sum_{i=1}^{M} (\boldsymbol{I}_N - \boldsymbol{\theta}_i \boldsymbol{\theta}_i^T)\,, \quad \boldsymbol{b} = \sum_{i=1}^{M} \boldsymbol{s}_i \quad \text{and} \quad c = \sum_{i=1}^{M} \boldsymbol{s}_i^T \boldsymbol{s}_i\,. \tag{A.7}$$

It is proven for equations with the form of eq. A.6 that the point $\boldsymbol{x}$ which minimizes eq. A.5 can be computed by

$$\boldsymbol{x} = \boldsymbol{A}^+ \boldsymbol{b} \tag{A.8}$$

with $\boldsymbol{A}^+$, the Moore-Penrose pseudo-inverse.
In cases of a (nearly) singular matrix $\boldsymbol{A}$, other solving methods, for example the singular value decomposition (SVD), have to be applied.

Derivation of the Projection Matrix The minimal distance from an arbitrary point x in $\mathbb{R}^3$ to a line $\boldsymbol{L}$ can be found by shifting the point x by $\boldsymbol{\theta}$ λ times along the line $\boldsymbol{L}$ until its position vector $\boldsymbol{x}$ is orthogonal to the direction vector $\boldsymbol{\theta}$ of $\boldsymbol{L}$. This mathematical operation can be expressed through

$$O(\boldsymbol{x}) = \boldsymbol{x} + \lambda\boldsymbol{\theta}. \tag{A.9}$$

which has to fulfill the orthogonality criterion

$$\langle O(\boldsymbol{x}), \boldsymbol{\theta}\rangle \stackrel{!}{=} 0. \tag{A.10}$$

By augmenting eq. A.10 with eq. A.9, we get

$$\langle \boldsymbol{x} + \lambda\boldsymbol{\theta}, \boldsymbol{\theta}\rangle = \langle \boldsymbol{x}, \boldsymbol{\theta}\rangle + \lambda\langle \boldsymbol{\theta}, \boldsymbol{\theta}\rangle \stackrel{!}{=} 0. \tag{A.11}$$

The transformation of eq. A.11 yields

$$\lambda = -\frac{\langle \boldsymbol{x}, \boldsymbol{\theta}\rangle}{\langle \boldsymbol{\theta}, \boldsymbol{\theta}\rangle}. \tag{A.12}$$

Inserting the definition of λ from eq. A.12 in eq. A.9 yields

$$\begin{aligned} O(\boldsymbol{x}) &= \boldsymbol{x} - \frac{\langle \boldsymbol{x}, \boldsymbol{\theta}\rangle}{\langle \boldsymbol{\theta}, \boldsymbol{\theta}\rangle}\boldsymbol{\theta} && \text{(A.13)}\\ &= \boldsymbol{x} - \langle \boldsymbol{x}, \boldsymbol{\theta}\rangle\boldsymbol{\theta} && \text{(A.14)}\\ &= \boldsymbol{x} - \boldsymbol{\theta}\langle \boldsymbol{x}, \boldsymbol{\theta}\rangle && \text{(A.15)}\\ &= \boldsymbol{x} - \boldsymbol{\theta}\boldsymbol{\theta}^T\boldsymbol{x} && \text{(A.16)}\\ O(\boldsymbol{x}) &= (\boldsymbol{I}_N - \boldsymbol{\theta}\boldsymbol{\theta}^T)\boldsymbol{x}. && \text{(A.17)} \end{aligned}$$

with the projection matrix $\boldsymbol{M} = \boldsymbol{I}_N - \boldsymbol{\theta}\boldsymbol{\theta}^T$.

A.3. Additional Figures for the LF_m Error Estimation

Fig. A.2 shows the error e_{LF_m} caused by a localization mismatches (azimuthal offsets) at the different microphone positions 1 to 30. Fig. A.3 depicts the error e_{LF_m} for the single microphone position 17 in the center and azimuthal offsets for loudspeakers 1-12 in the horizontal plane. It is observable that the e_{LF_m} for loudspeaker 1 belongs to the group of higher errors. Only positions in the corners (loudspeaker 6 and 11) and in the rear (loudspeaker 7) exhibit higher values in particular intervals. The highest maximum by far can be detected at loudspeaker 3. Apparently, a loudspeaker positioned in a corner of the room causes the arrival of an high amount of reflected sound energy at the Y-Channel of the microphone.
At microphone positions closer to the walls of the IEM CUBE, early reflections contribute strongly to the energy content of the $\tilde{Y}$-channel. Microphone position 17 is very little under the influence of the early reflections, since it is located farthest to the wall. Therefore, the increase in energy through the incoming direct sound due to a inaccurate rotation yields a rather big error e_{LF_m} and difference in the energy content of the $\tilde{Y}$-channel. Only the error at microphone position 23 is stronger influenced by a localization mismatch for loudspeaker 1 (see fig A.2).
The maximum of the most errors is typically located around 90° azimuthal offset, since there the figure-of-eight pattern points with its maximum sensitivity towards the sound source.

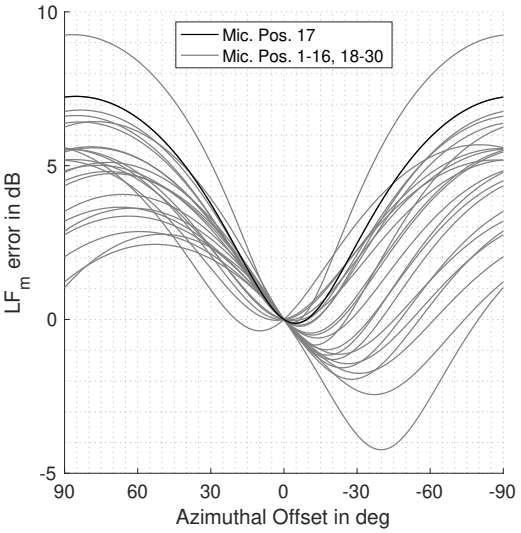


Figure A.2.: e_{LF_m} for LSP 1 and localization matches at mic. pos. 1 to 30.

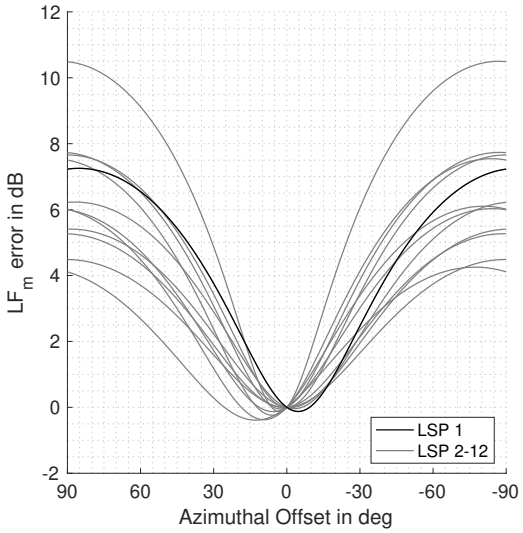


Figure A.3.: e_{LF_m} error for all loudspeakers from the central mic. pos. 17.

A.4. Additional Information to the Evaluations in Chapter 4

A.4.1. Additional Figures to the Level Equalization

György-Ligeti-Hall In the György-Ligeti-Hall, high level differences were expected, due to the large differences in distance of the individual loudspeakers to the center. The magnitude of the sound pressure p is proportional to the distance $\frac{1}{r}$. Thus, the relative distance factor F for every loudspeaker l is calculated with

$$F_l = \frac{r_l}{r_{\min}} \tag{A.18}$$

and multiplied with the belonging impulse responses $s(t)$ in order to compare the levels of the different loudspeakers:

$$s_{\mathrm{eq},l}(t) = F_l s_l(t) \; . \tag{A.19}$$

The level differences range in between $\pm$ 2.5 dB. Due to the large dimensions of the space and the loudspeaker setup, these deviations are tolerated and not equalized for the evaluation.

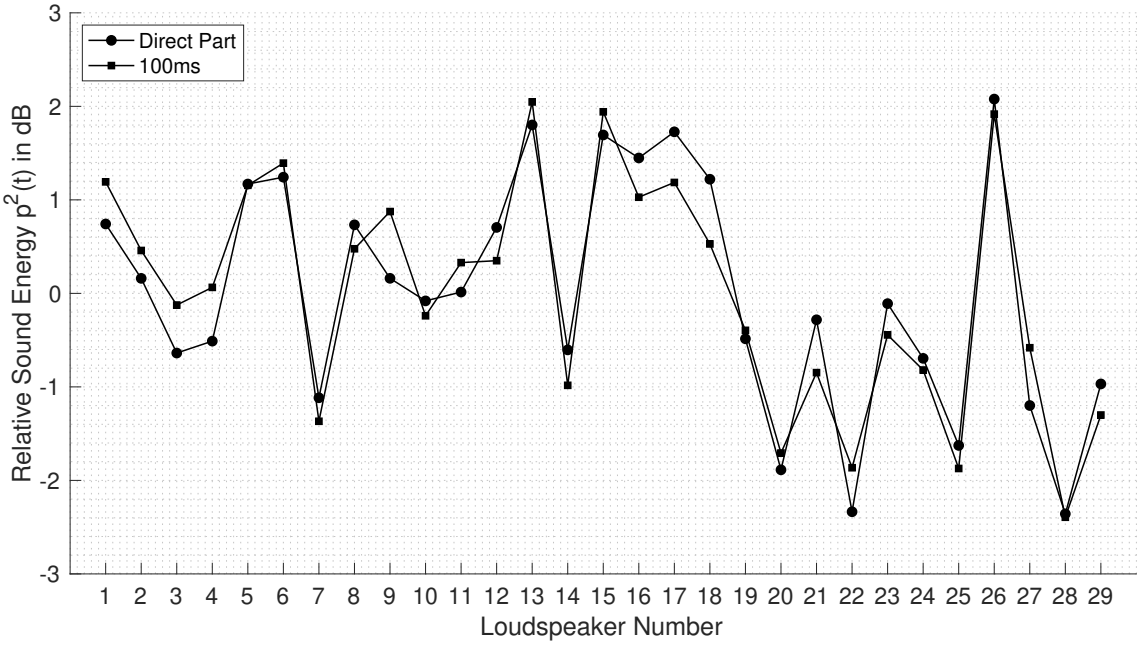


Figure A.4.: Sound energy in the direct part and 100 ms-part of the loudspeaker impulse responses in the György-Ligeti-Hall.

Production Studio In the Production Studio, loudspeakers 8 to 12 are too dominant if they are adjusted to the same gain as the lower ring of loudspeakers. Therefore, the upper ring is 3 dB SPL quieter and loudspeaker 12, is damped by 5 dB SPL.
In order to assess the surround sound system under the real use conditions, no gain compensation is applied.
Fig A.5 shows the deviations of the sound energy of the direct part and the first 100 ms of the impulse responses to the ideal sound energy, which would correspond to the predefined gains (LSP 1-7: 0 dB, LSP 8-11: -3 dB, LSP 12: -5 dB). The discrepancies of the direct part energy range between -0.8 and 1.0 dB. The energetic differences in the first 100 ms of the impulse response are smaller: The maximum can be found at loudspeaker 11 with 0.2 dB and loudspeaker 4 exhibits the minimal sound energy difference of -0.6 dB.

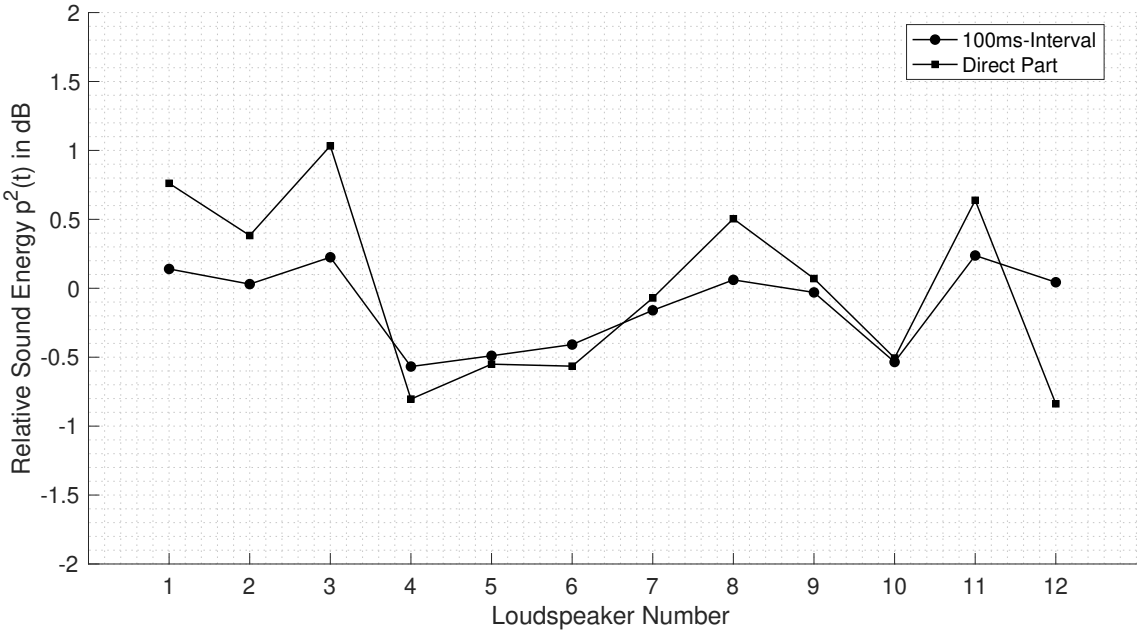
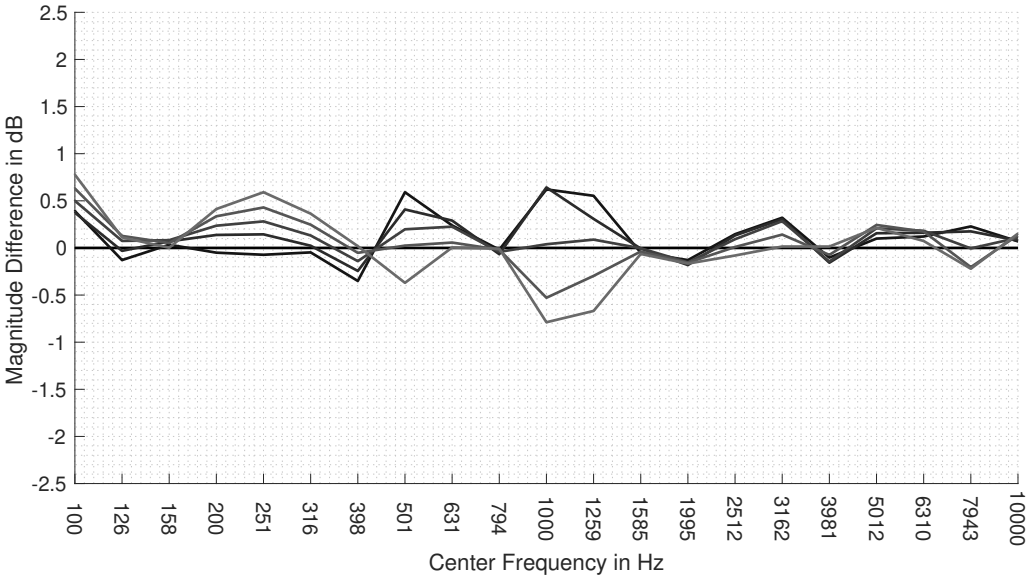


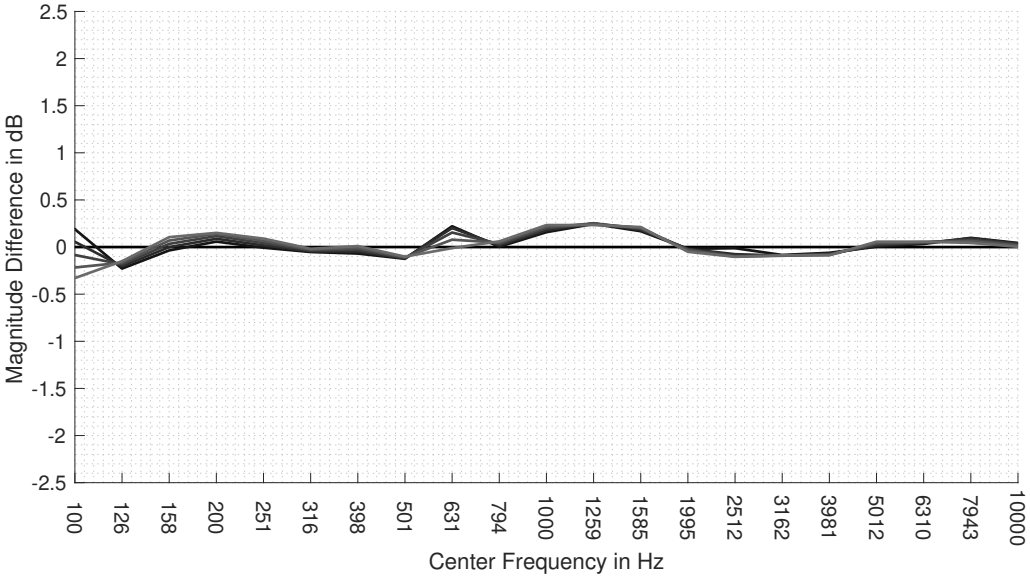
Figure A.5.: Sound energy of the direct part and the first 100 ms of the loudspeaker impulse responses in the Production Studio.

A.4.2. Additional Figures to the Coloration Measure (C_m)

Fig. A.6 shows the spectral difference plots of a phantom source, which is decoded with AllRAD and basic weighting in 1st- and 3rd-order Ambisonics. As before in the evaluation in the IEM CUBE, the phantom source is decoded in six steps of 5 ° from loudspeaker 1 to loudspeaker 2.



(a) 3rd-order Ambisonics.



(b) 1st-order Ambisonics.

Figure A.6.: Spectral difference plots for third-octave bands using basic-weighted AllRAD.