Quality Assessment of Spherical Microphone Array Auralizations

Dissertation zur Erlangung des
akademischen Grades Doktor-Ingenieur (Dr.-Ing.)

vorgelegt der Fakultät für Elektrotechnik und Informationstechnik
der Technischen Universität Ilmenau
von Dipl.-Ing. Johannes Nowak

1. Gutachter: Prof. Dr.-Ing. Dr. rer. nat. h.c. mult. Karlheinz Brandenburg
2. Gutachter: Prof. Dr.-Ing. Boaz Rafaely
3. Gutachter: Dr.-Ing. Frank Melchior

Tag der Einreichung: 07.02.2019
Tag der wissenschaftlichen Aussprache: 24.06.2019

urn:nbn:de:gbv:ilm1-2019000222

Technische Universität Ilmenau

Fraunhofer Institute for Digital Media Technology IDMT

Doctoral Thesis

Quality Assessment of Spherical Microphone Array Auralizations

by

Johannes Nowak

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Engineering

in the

Electronic Media Technology Laboratory
Technische Universität Ilmenau

July 2, 2019

“Trying to predict the future is a mug’s game. But increasingly it’s a game we all have to play because the world is changing so fast and we need to have some sort of idea of what the future’s actually going to be like because we are going to have to live there, probably next week."

Douglas Adams
(1952–2001)

Acknowledgements

This thesis would not have been possible without the motivation and support from the following people; I'd like to thank

...Prof. Karlheinz Brandenburg, my advisor, for giving me the opportunity to work in his research team at Fraunhofer IDMT/TU Ilmenau, his guidance and motivation as well as for his critical view and objective advices;

...Prof. Boaz Rafaely for his technical insight and support, his analytical perspective, and for pointing me (and this thesis) in the right direction;

...Dr. Frank Melchior for his encouragement and inspiration, fruitful discussions, and stimulating off-topic talks;

...my colleagues at Fraunhofer IDMT and TU Ilmenau, namely Dr. Sandra Brix, Prof. Thomas Sporer, Dr. Daniel Beer, Andreas Männchen, Rene Rodigast, Christoph Sladeczek, Michael Strauß, Clemens Clausen, Thomas Korn, Claudia Heinze, Judith Liebetrau, Lutz Ehrig, Javier Frutos-Bonilla, Alejandro Gazul-Ruiz, Gyorgy Nagy, Tobias Claus, Mario Seideck, Tobias Gehlhaar, Sven Kämpf, Michael Rath, Albert Zykhar, Wolfgang Lorenz, Annika Neidhardt, Bernhard Fiedler, Matthias Fiedler, Paul Fritzsche, Ania Lukashevich, Sascha Grollmisch, Patrick Aichroth, Andras Katai, Katrin Pursche, Mareike Helbig, Julia Hallebach, Ulrike Sloma, Thomas Köllmer, Stephan Werner, Florian Klein, Sara Kepplinger, Frank Hofmeier, Dr. Eckhardt Schön, Thaden Cohrs, and Andre Siegel as well as Prof. Gerald Schuller, Prof. Alexander Raake, and Prof. Hans-Peter Schade; Also, I'd like to thank the staff of Fraunhofer IDMT and TU Ilmenau.

...my colleagues of the SEACEN research unit, namely Prof. Benny Bernschütz, Dr. Alexander Lindau, Vera Erbes, Fabian Brinkmann, Mathias Geier, Frank Schultz, Stefan Klockgether, Philip Stade as well as Prof. Sascha Spors, Prof. Steven Van De Par, Prof. Christoph Pörschmann, Prof. Stefan Weinzierl, and Prof. Michael Vorländer;

...all my students, specifically Viktor Böhm, Alexander Albrecht, Sascha Pälchen, Martina Hockardt, Josua Hagedorn, Jan Küller, Timo Gabb, Robin Ritter, Lorenz Betz, Maximilian Wolf, Mathias Hellmich, Anna Rüppel, Georg Fischer, Tobias Brocks, and Kai-Peter Jurgeit;

...Dr. Andreas Franck, Judith Liebetrau, Stephan Jaroschek, and Toini Schmitz for proofreading;

...my friends and family—you know who you are;

...and my girlfriend Sylke for her love.

Abstract

The thesis documents a scientific study on quality assessment and quality prediction in Virtual Acoustic Environments (VAEs) based on spherical microphone array data, using binaural synthesis for reproduction. In the experiments, predictive modeling is applied to estimate the influence of the array on the reproduction quality by relating the data derived in perceptual experiments to the output of an auditory model.

The experiments adress various aspects of the array considered relevant in auralization applications: the influence of system errors as well as the influence of the array configuration employed. The system errors comprise spatial aliasing, measurement noise, and microphone positioning errors while the array configuration is represented by the sound field order in terms of spherical harmonics, defining the spatial resolution of the array. Based on array simulations, the experimental data comprise free-field sound fields and two shoe-box shaped rooms, one with weak and another with strong reverberation. Ten audio signals served as test material, e.g., orchestral/pop music, male/female singing voice or single instruments such as castanets.

In the perceptual experiments, quantitative methods are used to evaluate the impact of system errors while a descriptive analysis assesses the array configuration using two quality factors for attribution: Apparent Source Width (ASW) and Listener Envelopment (LEV). Both are quality measures commonly used in concert hall acoustics to describe the spaciousness of a room. The results from the perceptual experiments are subsequently related to the technical data derived from the auditory model in order to build, train, and evaluate a variety of predictive models. Based on classification and regression approaches, these models are applied and investigated for automated quality assessment in order to identify and categorize system errors as well as to estimate their perceptual strength. Moreover, the models allow to predict the array's influence on ASW and LEV perception and enable the classification of further sound field characteristics, like the reflection properties of the simulated room or the sound field order used. The applied prediction models comprise simple linear regression and decision trees, or more complex models such as support vector machines or artificial neural networks.

The results show that the developed prediction models perform well in their classification and regression tasks. Although their functionality is limited to the conditions underlying the conducted experiments, they can still provide a useful tool to assess basic quality-related aspects which are important when developing spherical microphone arrays for auralization applications.

Zusammenfassung

Die vorliegende Arbeit beschäftigt sich mit der Qualitätsbewertung und -vorhersage in virtuellen akustischen Umgebungen, insbesondere in Raumsimulationen basierend auf Kugelarraydaten, die mithilfe binauraler Synthese auralisiert werden. Dabei werden verschiedene Prädiktionsverfahren angewandt, um den Einfluss des Arrays auf die Wiedergabequalität automatisiert vorherzusagen, indem die Daten von Hörexperimenten mit denen eines auditorischen Modells in Bezug gesetzt werden.

Im Fokus der Experimente stehen unterschiedliche, praxisrelevante Aspekte des Messsystems, die einen Einfluss auf die Wiedergabequalität haben. Konkret sind dies Messfehler, wie räumliches Aliasing, Rauschen oder Mikrofonpositionierungsfehler, oder die Konfiguration des Arrays. Diese definiert das räumliche Auflösungsvermögen und entspricht der gewählten Ordnung der Sphärischen Harmonischen Zerlegung. Die Experimente basieren auf Kugelarray-Simulationen unter Freifeldbedingungen und in einfachen simulierten Rechteckräumen mit unterschiedlichen Reflexionseigenschaften, wobei ein Raum *trocken*, der andere dagegen stark *reflektierend* ist. Dabei dienen zehn Testsignale als Audiomaterial, die in praktischen Anwendungen relevant erscheinen, wie z. B. Orchester- oder Popmusik, männlicher und weiblicher Gesang oder Kastagnetten.

In Wahrnehmungsexperimenten wird der Einfluss von Messfehlern in einer quantitativen Analyse bewertet und die Qualität der Synthese deskriptiv mit den Attributen Apparent Source Width (ASW) und Listener Envelopment (LEV) bewertet. Die resultierenden Daten bilden die Basis für die Qualitätsvorhersage, wobei die Hörtestergebnisse als Observationen und die Ausgangsdaten des auditorischen Modells als Prädiktoren dienen. Mit den Daten werden unterschiedliche Prädiktionsmodelle trainiert und deren Vorhersagegenauigkeit anschließend bewertet. Die entwickelten Modelle ermöglichen es, sowohl Messfehler zu identifizieren und zu klassifizieren als auch deren Ausprägung zu schätzen. Darüberhinaus erlauben sie es, den Einfluss der Arraykonfiguration auf die Wahrnehmung von ASW und LEV vorherzusagen und die verwendete Ordnung der Schallfeldzerlegung zu identifizieren, ebenso wie die Reflexionseigenschaften des simulierten Raumes. Es kommen sowohl einfache Regressionsmodelle und Entscheidungsbäume zur Anwendung als auch komplexere Modelle, wie Support Vector Machines oder neuronale Netze.

Die entwickelten Modelle zeigen in der Regel eine hohe Genauigkeit bei der Qualitätsvorhersage und erlauben so die Analyse von grundlegenden Array-Eigenschaften, ohne aufwendige Hörexperimente durchführen zu müssen. Obwohl die Anwendbarkeit der Modelle auf die hier untersuchten Fälle beschränkt ist, können sie sich als hilfreiche Werkzeuge bei der Entwicklung von Kugelarrays für Auralisationsanwendungen erweisen.

Contents

Acknowledgements ii

Abstract iii

Zusammenfassung iv

1 Introduction 1

1.1 Audio Quality Assessment . . . 2

1.2 Virtual Acoustic Environments . . . 3

1.2.1 Room Simulation based on Binaural Technology . . . 4

1.2.2 Room Simulation based on Microphone Array Data . . . 6

1.3 Outline of this Thesis . . . 8

1.3.1 Research Questions . . . 10

1.3.2 Contribution . . . 11

1.4 Organization of Contents . . . 11

2 State of Research: Audio Quality & Quality Prediction 13

2.1 Quality of Concert Halls . . . 13

2.2 Quality of Virtual Acoustic Environments . . . 14

2.3 Quality of Microphone Array Auralizations . . . 15

2.4 Quality Prediction . . . 16

2.4.1 Quality Prediction in Audio Coding . . . 17

2.4.2 Quality Prediction in Concert Hall Acoustics . . . 18

2.4.3 Quality Prediction in Spatial Audio . . . 19

3 Theoretical Background 20

3.1 Spherical Sound Fields . . . 20

3.1.1 Spherical Harmonics . . . 21

3.1.2 Fundamental Waves in Spherical Coordinates . . . 23

3.1.3 Sampling Order-Limited Functions . . . 26

3.1.4 Spherical Array Beamforming 34
3.1.5 Measurement Errors 39
3.1.6 Head-Related Transfer Functions 44
3.2 Perceptual Fundamentals 45
3.2.1 Human Auditory System 45
3.2.2 Binaural Hearing 53
3.2.3 Perception of Spaciousness 55
3.2.4 Auditory Models 58
3.3 Predictive Modeling 64
3.3.1 Common Predictive Modeling Techniques 64
3.3.2 Model Performance Evaluation 72
3.3.3 Data Preprocessing 73
3.3.4 Model Training and Tuning 74
3.3.5 Predictor Importance 76
3.3.6 Problems in Predictive Modeling 77

4 Perceptual Evaluation **79**
4.1 Experiment I: Quantitative Error Analysis under Free-Field Conditions . 82
4.1.1 Listening Test Design 83
4.1.2 Listening Test Results 83
4.1.3 Discussion 87
4.2 Experiment II: Quantitative Error Analysis in Reflective Environments . 87
4.2.1 Room Simulations 88
4.2.2 Listening Test Design 89
4.2.3 Listening Test Results 89
4.2.4 Discussion 93
4.3 Experiment III: Descriptive Quality Analysis based on ASW and LEV . . 94
4.3.1 Listening Test Design 94
4.3.2 Listening Test Results 95
4.3.3 Discussion 98

5 Predictive Modeling **100**
5.1 Error Prediction 102
5.1.1 Classification in Free-Field Environments 106
5.1.2 Regression in Free-Field Environments 112
5.1.3 Classification in Reverberant Environments 118

5.1.4 Regression in Reverberant Environments 120
5.1.5 Experiment: Error Comparison between Rooms 122
5.2 Quality Prediction 126
5.2.1 Regression Analysis for ASW and LEV Prediction 126
5.2.2 Predictor Importance 127
5.2.3 Discussion 128
5.3 Prediction of Microphone Array Configurations and Sound Field Characteristics 129
5.3.1 Classification of Microphone Array Configurations 129
5.3.2 Classification of Reflection Properties 131
5.3.3 Predictor Importance 132
5.3.4 Discussion 132

6 Conclusions 133

6.1 Perceptual Evaluation 133
6.2 Predictive Modeling 135
6.3 Limitations 136

7 Future Work 138

List of Figures 140

List of Tables 142

List of Acronyms 144

Bibliography 149

A Room Acoustical Fundamentals 177

A.1 Room Impulse Response 177
A.2 Room Acoustical Parameters 178

B Spherical Sound Field Special Functions 180

B.1 Wave Equation in Spherical Coordinates 180
B.2 Spherical Base Solutions 183
B.3 Spherical Fourier Transform 189
B.4 Spherical Convolution and Correlation 192

C Tables 193

D Perceptual Data 206

Eigenständigkeitserklärung

1 Introduction

Today, the quality of Virtual Reality (VR) applications has reached a level of realism that, not long ago, was thought of as science fiction. Although the *Holodeck* of Star Trek is still a vision, some recent developments in multimedia technology allow a glimpse of what VR might look like in the near future. By means of digital signal processing, it is already possible to design and render complex virtual worlds which can be reproduced using technologies such as VR glasses or spatial audio playback. The synthesized virtual environments provide users with an interactive and immersive multimedia experience by recreating lifelike sensory impressions. In practice, however, a variety of technical, physical, and signal-processing-related limitations—as well as cognitive aspects such as the users' expectations—affect the quality of the immersion.

Quality, or Quality of Experience (QoE), is essential for multimedia products as it directly influences their acceptance and, consequently, their market success. QoE is defined as the "degree to which a set of inherent characteristics fulfills requirements" [3] or the "perception of the degree to which [costumer or user] requirements have been fulfilled" [4]. To meet these requirements, it is necessary to assess the reproduction quality using perceptual experiments which, in practice, are often time consuming and expensive. In order to overcome these detriments, predictive modeling can be applied to predict the human response to certain stimuli [14], hence, predicting quality. Quality predicion relies on technical features extracted from digital models of human perception, and statistical models relating these features to the results of perceptual experiments.

This thesis exclusively investigates *acoustic* VRs based on binaural auralizations of 3D sound fields recorded with a spherical microphone array, specifically focussing on how system errors and the array configuration impair the reproduction quality. The quality assessment is done perceptually by assessing the reproduction quality in listening experiments, and technically by applying a model of human auditory perception. The resulting data is subsequently combined employing a variety of prediction models. Thereby, a tool for automated quality prediction is provided when designing spherical microphone arrays for room simulation applications.

1.1 Audio Quality Assessment

Quality is a multimodal percept, including sensorial and higher-level cognitive aspects such as knowledge, emotions, or experience [188, 103]. Although all human senses affect the perception of VRs in practical applications, this thesis solely focusses on the assessment of the acoustic auralization quality. Accordingly, the following descriptions address different aspects of audio quality and its assessment. Recently published overviews regarding audio quality assessment can be found, for example, in [53, 265].

The assessment of audio quality is based on human judgments which can be measured in formal listening experiments. The response of a listener to an audio stimulus provides a quantitative measure for any auditory impression of interest. Then, the measured data can be related to the technical or acoustical domain for further system optimization. According to [14], three different approaches for audio quality assessment exist:

- **Quantitative assessment** evaluates basic audio quality, providing a measure for an overall auditory impression by assessing attributes like naturalness or plausibility. Quantitative assessment is common in standardized acoustical quality measurement procedures, like in recommendation ITU-R BS.1284 [51].

- **Qualitative (or descriptive) assessment** refers to the sensorial strength of an individual auditory attribute. This kind of assessment provides more information on specific perceptual aspects based on individual quality factors, like attributes derived from an interview-based Free-Choice Profiling (FCP) approach [298] or using consensus [306] or individual vocabulary profiling [125, 74, 160]. Combining quantitative and qualitative assessment, a mixed-method approach was proposed in [270, 269], the so-called Open Profiling of Quality (OPQ).

- **Predictive modeling** allows the estimation of a listener's judgment regarding reproduction quality. This approach is used, for example, in audio coding quality assessment [276]. The modeling process comprises two steps: first, an auditory model evaluates specific properties of a stimulus providing technical/physical data; second, a subsequent statistical model links perceptual and technical data (e.g., data derived from an impulse response) for quality prediction.

These three approaches of audio quality assessment are fundamental to this thesis, as synthesized sound fields are to be evaluated by means of both quantitative and qualitative assessment, producing data which is then used for predictive modeling. An overview of the state of research in acoustic quality assessment is provided in Chapter 2.

1.2 Virtual Acoustic Environments

In acoustics, a VR is commonly referred to as a Virtual Acoustic Environment (VAE) [241]. A VAE is a virtual sound space which comprises complex acoustical scenes, containing foreground and background objects which can be modeled by multiple, spatially distributed virtual sound sources. The VAE data can be rendered for audio playback by means of digital signal processing and reproduced using loudspeaker arrays or headphones, like in Wave Field Synthesis (WFS) [26] and Higher-Order Ambisonics (HOA) [165, 72], or binaural synthesis [177, 190, 156], respectively. A recent review on loudspeaker array-based approaches, including perceptual aspects, can be found in [266, 232]. Figure 1.1 shows the process of implementing a VAE.

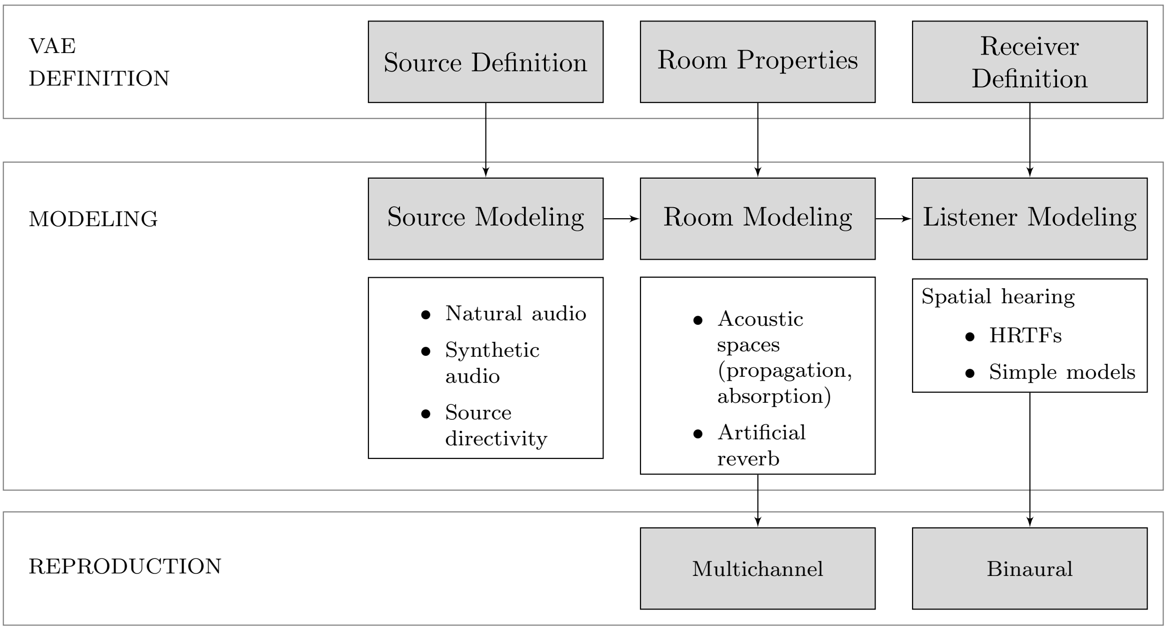


Figure 1.1: Implementation of a VAE based on three stages: definition, modeling, and reproduction (after [241]).

It comprises three stages: definition, modeling, and reproduction. The term *virtual acoustics* primarily refers to the second stage which includes the modeling of the sound source's properties, like its directivity [308], the room characteristics, and the (spatial) hearing properties of the listener which are commonly modeled using Head-Related Transfer Functions (HRTF). The second stage (modeling) and the third stage (reproduction) together are known as *auralization* [136]. In practice, the data that describes a VAE can be synthesized or measured, like with a microphone array [123] or an artificial head [156]. A comprehensive overview of auralization and VAEs is given in [18, 286, 232].

1.2.1 Room Simulation based on Binaural Technology

Binaural synthesis can be realized by placing microphones at the ear drums of a listener (or dummy head) and, in order to provide the listener with a realistic impression of the recorded room, playing back the recorded signals via headphones [177, 190]. In practice, however, the sole reconstruction of the sound pressures at the ear drums is insufficient as cognition strongly affects the reproduction quality [278, 207, 129]. If the listener's expectations cannot be met, then the synthesis quality collapses [60, 295, 42].

An important quality metric for binaural synthesis is externality, a descriptive attribute that describes the degree to which the listener locates an acoustical event outside of his head (The opposite effect is the so-called *in-head localization* [146]). Externality is directly related to the immersive experience known from VR systems in a way that the immersion is substantially reduced when externality is low, and vice versa [286].

Binaural recording: For binaural recording of, for instance, a concert hall, a pair of microphones is placed at the ear drums of a human listener for *individual* (or of an artificial head for *non-individual*) binaural synthesis [71, 152]. For practical reasons, the recording at the blocked ear canal is sufficient, since sound propagation from the ear canal entrance to the ear drum is independent of the sound arrival direction [172, 112]. Commonly, the data recorded is stored as a set of Binaural Room Impulse Responses (BRIR) comprising the binaural transfer paths from a sound source in a room to the receiver. Recent improvements include Head-and-Torso Simulators (HATS), allowing to record a sound field for different head orientations [176, 163, 158, 119]. Recordings can be rendered for Dynamic Binaural Synthesis (DBS), i.e., including head-tracking, which basically marks the beginning of binaural synthesis as it is understood today.

Binaural reproduction: Recent developments are based on Head-Related Transfer Functions (HRTF)[1]. Comprising the transfer characteristics of torso, head, and pinnae, a HRTF describes the free-field transfer path from a sound source to both ear drums. It is represented by a two-channel audio signal which can be implemented as a filter with a Finite Impulse Response (FIR). HRTFs are a flexible tool for binaural resynthesis, allowing the simulation of various environments, which can be data-based using array recordings, or model-based using a room simulation software. In practical applications, binaural synthesis is commonly implemented using headphones, but also loudspeaker-based binaural reproduction is possible [70, 284].

[1] Note that a single HRTF comprises a signal pair for one sound incident direction. The plural, HRTFs, describes a set of HRTFs for multiple directions. Single-channel HRTFs are denoted accordingly.

Limitations: The fundamental problem in binaural technology lies in the anthropometry of the listener's torso, head, and pinnae. If it significantly differs to the one of the HRTFs used, then the binaural reproduction quality is dramatically decreased. In addition, various issues are known to affect the reproduction quality in practical applications. An overview on some prominent issues is given in the following:

- Front-back confusion is a prominent issue. It describes the effect that a sound source presented in front of the listener is perceived as coming from behind [179, 180, 175]. Applying head-tracking in the reproduction system can help minimize this effect [156].

- Experiments show that the listener's natural head movements are important for spatial perception and should be included into the synthesis to increase the reproduction quality in terms of externality [293, 163]. In practice, this is achieved using head-tracking, i.e., DBS.

- When using head-tracking, the method applied to interpolate between different HRTFs also has an audible effect on the reproduction [49].

- Also, the type of headphone used for playback influences the synthesis quality, gradually increasing from closed to half-open, to open headphones [243]. Best results can be achieved with extra-aural headphones [249, 80].

- Equalization of the headphone's transfer characteristics further improves the synthesis quality [243].

- When recording HRTFs/BRIRs, both the environment and the measurement procedure have an effect on the synthesis quality [285].

- Context dependent parameters strongly influence the immersion [294], such as differing acoustical properties between the listening and the synthesized room. In this case, adjusting the direct-to-reverberation ratio can optimize the reproduction quality [296].

- Investigations in [135] show that training effects and long-term customization can improve the quality of binaural synthesis.

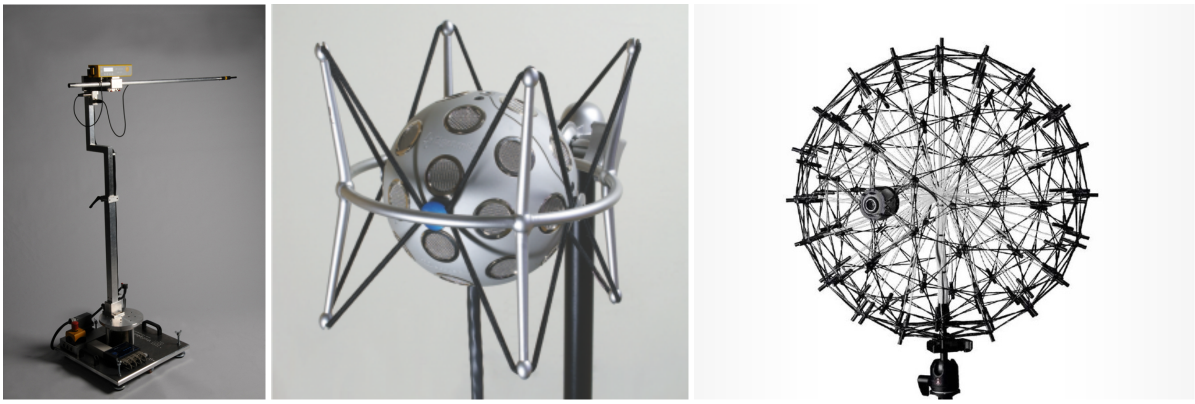
It should be noted that only a few of the listed criteria for optimal reproduction could be met in the experiments conducted in Chapter 4. Consequently, the results presented are only valid within their experimental conditions. However, some findings may also be generalized for other reproduction system setups.

1.2.2 Room Simulation based on Microphone Array Data

Microphone arrays provide another, more flexible way for spatial sound field recording. With the geometrical arrangement of multiple microphones working in tandem, it is possible to sample a sound field spatially, thereby recording its directional properties. Once the sound field is recorded (and stored), the data can be rendered and synthesized on any reproduction system. Moreover, beamforming can be applied for spatial analysis. In [167], methods for spatial sound design are proposed for artistic purposes.

Early microphone array configurations were basically multi-channel setups for stereo, 5.1-, or 7.1-recordings. In the following years, other geometries, like linear, circular, or spherical arrays were employed for auralization applications [123], offering new possibilities for sound field processing. An early 3D configuration was the so-called *sound-field microphone* which was employed for ambisonic recordings in the 1970's [101]. Recent developments include high-resolution spherical arrays for room recordings [171, 225, 167], or the directivity characteristic analysis of musical instruments [308].

Array recording: Basically two approaches for array recordings exist: The first drives multiple microphones in tandem while the second employs a robot arm with a single microphone, successively sampling the sound field. These so-called virtual (or scanning) microphone arrays offer high flexibility in terms of sensor positioning and spatial resolution, but lack real-time applicability. In practice, the recorded sound field is typically stored as a set of directional Room Impulse Responses (RIR) [143, 84]. Figure 1.2 exemplarily shows some spherical microphone arrays.



(a) VariSphere [30] (b) Eigenmike [126] (c) Acoustic Camera [102]

Figure 1.2: Examples of spherical microphone arrays: a) single-channel scanning array, b) 32-channel rigid-sphere array, and c) an 120-channel open-sphere array.

Array data reproduction: Once the sound field is recorded and made available for auralization, it can be rendered on any playback system by means of Plane Wave Decomposition (PWD) [171, 217, 220, 224]. This approach is used, for example, in stereophonic or WFS playback [123, 167] as well as in binaural synthesis [8, 193]. Binaural auralization of spherical microphone array data is realized convolving a high-resolution spherical set of HRTFs with the array recordings, resulting in a pair of BRIRs which comprise the three-dimensional acoustic wave field.

In spherical microphone array applications, the sound field is usually represented in the spherical harmonics domain. Spherical harmonics are a set of basis functions forming the Fourier basis for functions on a sphere. They constitute a flexible and intuitive sound field representation and are fundamental for microphone array data processing in spherical coordinates (see Chapter 3.1 for details). Order N of the spherical harmonic decomposition represents the spatial accuracy of the microphone array, which, in practice, is limited by the array configuration used.

Limitations: In practice, the operational bandwidth of the microphone array is limited by the properties of the array configuration, the sampling scheme used, and measurement errors that occur during the recording process [215]. Errors corrupt the spatial response of the array, leading to degradations in the sound field representation. This is especially critical in auralization applications for which a full audio bandwidth is desired. Such errors are:

- spatial aliasing [215, 222] which distorts the array response at high frequencies, thereby introducing annoying high-frequency ringing sounds [8].
- uncorrelated noise, e.g. microphone noise, results in extensive low frequency amplifications due to numerical ill-conditionings (division by zero) when performing a PWD at high orders [215].
- positioning errors can corrupt the auralization quality [215] as the PWD relies on accurate sample positions on the measurement grid.
- non-ideal directivity characteristics also affect the accuracy of the PWD, consequently corrupting the auralization quality [76, 167].

In the scope of this thesis, the influence of spatial aliasing, microphone noise, and positioning errors on the auralization quality is investigated. A perceptual analysis of how the directivity of the employed microphones influences the reproduction quality is not addressed and proposed for future work.

1.3 Outline of this Thesis

The central outcome of this thesis is a *quality assessment toolbox* for automated evaluation of the reproduction quality when using spherical microphone arrays for room simulation applications. This toolbox aims at optimizing the array design process, as it allows to evaluate important aspects of the array—and their impact on the reproduction quality—without the need to conduct time-consuming listening experiments.

The presented investigations are based on a three-step procedure: first, the influence of the array configuration and system errors on the reproduction quality is evaluated based on three listening experiments; second, a technical analysis is conducted using an auditory model; third, both perceptual and technical data are used for predictive modeling, enabling the automated estimation of an assessor's quality judgments.

The experiments are based on simulated data, evaluating different array configurations in terms of varying sound field orders, but also addressing system errors such as spatial aliasing, measurement noise, and microphone positioning errors. The array simulations comprise sound fields under free-field conditions and of two shoe-box shaped rooms characterized by varying reflection properties (one room with low, the other with high reverberation). These array simulations are combined with a spherical set of HRTFs from a dummy head for binaural auralization. The resulting BRIRs are further convolved with various audio signals in order to also address the impact of the content characteristics on the reproduction quality. Specifically, the following listening experiments are conducted:

- **Experiment I** investigates the impact of measurement errors on the reproduction quality, assessing array simulations under free-field conditions in a quantitative analysis.

- **Experiment II** evaluates the impact of measurement errors in reflective environments based on a quantitative analysis.

- **Experiment III** is a descriptive analysis using two quality metrics commonly used in concert hall acoustics to describe reproduction quality: Apparent Source Width (ASW) and Listener Envelopment (LEV). Based on simulations in different reflective environments, the array configuration used and its impact on the reproduction quality is evaluated in terms of ASW and LEV.

As a result, multiple stimuli are available for further analysis, providing a suitable data set to develop models for quality prediction. The data resulting from the perceptual experiment, termed *observations* or *dependent variables*, is used to relate perception to

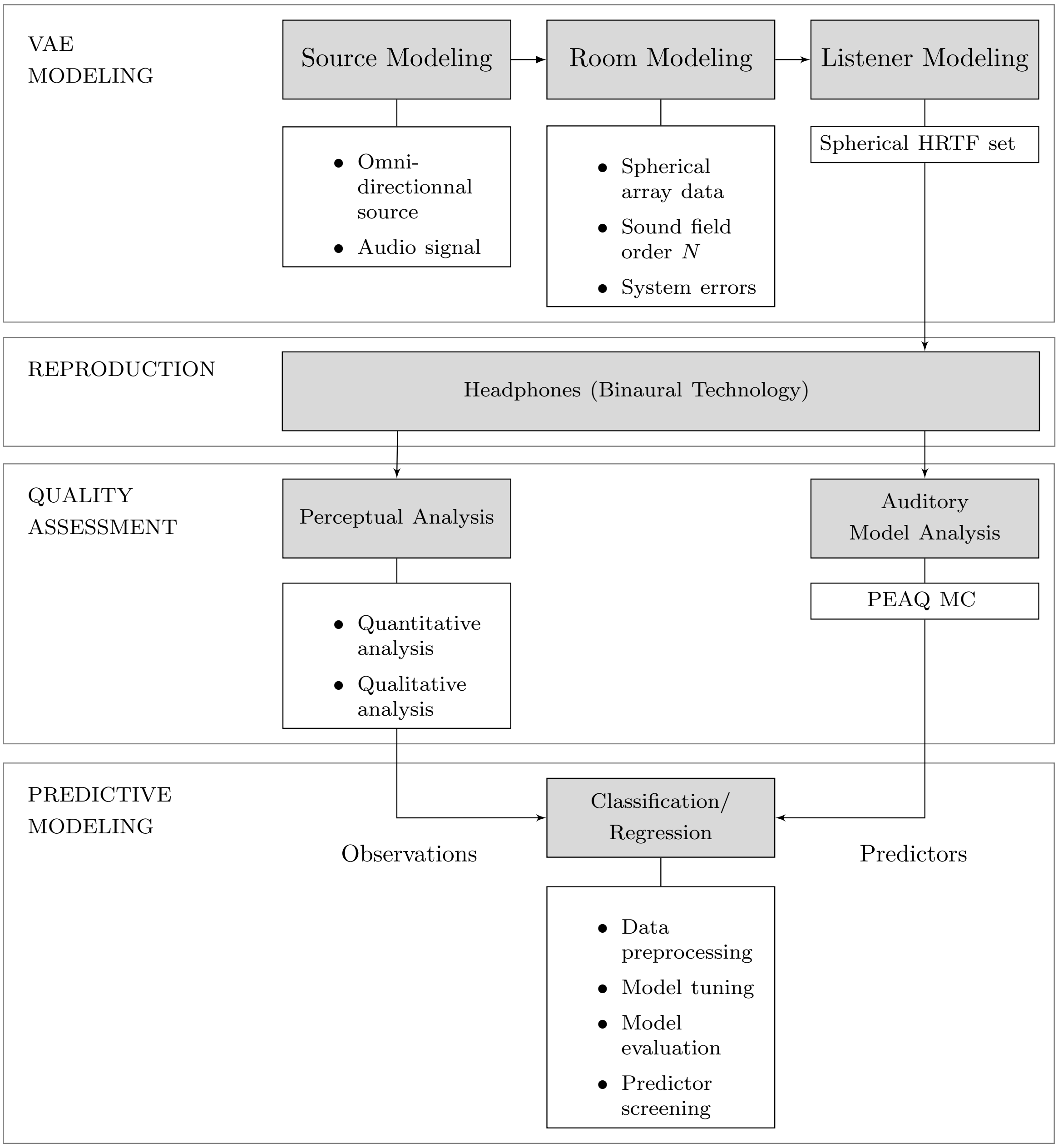


Figure 1.3: Workflow schematic fundamental to the analyses conducted in this thesis.

technical predictor, the so-called *predictors*, *features* or *independent variables*, based on which the developed model estimates the quality rankings from the assessors. These variables are technical measures represented by the output of an auditory model, hence, Model Output Variables (MOV). In particular, the multi-channel model of Perceptual Evaluation of Audio Quality (PEAQ-MC) is used for evaluation [276, 155]. It provides 39 MOVs comprising distortion and modulation measures as well as binaural (spatial) features, like interaural differences. Using these features to analyze the same audio data that was presented to the assessors, it is possible to establish a relation between technical and perceptual data. This is achieved with a variety of predictive modeling techniques, among them linear regression models, neural networks, support vector machines, and decision trees. Each model is trained with a portion of the data, tuned and generalized in terms of their predictive performance, and finally tested against the remaining samples to see how these models would perform on unknown data. The aim of prediction is to estimate the strength of an error and the quality in terms of ASW and LEV, all based on a regression analysis. In addition, a classification approach is used to identify and categorize measurement errors, microphone array configurations, and sound field characteristics.

Although predictive modeling has already been applied in various quality-related experiments, like in [99], [155] or [137], it was not yet used to assess spherical microphone arrays for auralization applications. In addition, when such techniques were applied for prediction, the models used were mainly simple linear models to link technical and perceptual data in a curve-fitting approach. In practice, such simple models are prone to miss important relationships within the data, especially when the quality feature of interest can only be modeled by a combination of predictors. More advanced models, like neural networks, can achieve such differentiations, usually providing better prediction performances and higher robustness against outliers or missing data.

1.3.1 Research Questions

The main research question underlying this thesis is formulated as follows:

> Is it possible to predict the reproduction quality of (binaural) spherical microphone array auralizations?

Based on this question, four supplementary research questions arise:

1. What is the perceptual effect of measurement errors on the auralization quality?
2. What is the influence of different test signals and reflection properties on the auralization quality?
3. What is the influence of the array configuration used on reproduction quality (in terms of ASW and LEV)?
4. How accurate is the prediction and can the results be generalized for new, unknown samples?

1.3.2 Contribution

In order to answer these research questions, this thesis contributes to the common state of research with:

- a comprehensive perceptual analysis of measurement errors and various array configurations based on quantitative and descriptive analysis methods, using free-field and simple room simulations with varying reflection properties (also taking different array configurations and multiple test signals into account);
- the technical evaluation of system errors and the array configuration used in terms of a predictor importance analysis, relating the parameters investigated to the MOVs of an auditory model;
- the development and evaluation of predictive models for error detection, classification, and quantification;
- the development and evaluation of predictive models for quality assessment in terms of ASW and LEV;
- the application of these models to classify array order and reflection properties based on the VAE data.

As an extension of this thesis, the author would like to reference a recent publication: The analysis presented in [194] deals with quality assessment and prediction of spherical array auralizations using ASW and LEV for attribution. In the experiment presented, the model described in [137] is used for prediction, providing low error scores and high correlations with the ratings from assessors.

1.4 Organization of Contents

The content of this thesis is structured as follows: This introductory chapter provided an overview of VAEs based on spherical microphone array data and binaural technology. Furthermore the outline of this thesis with the definition of research questions were given.

The following Chapter 2 presents the state of research related to audio quality and quality prediction. Particularly, the first section addresses audio quality assessment of concert halls, VAEs and, in particular, the evaluation of microphone array auralizations. The second section gives an overview on quality prediction methods used in audio coding research and concert hall acoustics.

The fundamentals important for this thesis are provided in Chapter 3, addressing the theoretical framework for sound field description in spherical coordinates. In this regard, an overview is given on spherical harmonics, sound field sampling with spherical arrays, and common array processing techniques such as like beamforming and sound field decomposition as well as an analytical description of measurement errors. The following Section 3.2 deals with fundamental psychoacoustic principles which are important in the scope of this thesis, like the human auditory system, the properties of spatial hearing, the impression of spaciousness, specifically the perception of ASW and LEV, and the description of the PEAQ model which is used in the experiments for technical analysis. Section 3.3 provides an overview of predictive modeling, addressing common models, evaluation methods, and problems as well as data pre- and postprocessing techniques, like resampling, performance analysis, and predictor importance evaluations.

Chapter 4 presents the conducted listening experiments[2]. In particular, measurement errors are analyzed in two experiments, where the first addresses errors under free-field conditions and the second in reflective environments. Then, the quality of different array configurations is assessed in Experiment III using ASW and LEV as descriptors.

Predictive modeling is applied in Chapter 5 in a threefold analysis which is based on the data from the listening tests and the output from the auditory model. The first part deals with system errors which are identified in a classification task and assessed in their strength based on a regression analysis. Second, the reproduction quality is predicted in terms of ASW and LEV using regression models while the third part is a classification approach to assess further sound field characteristics, like the array configuration used and the reflection properties of the simulated environment.

Finally, Chapter 6 summarizes the presented work, taking the experimental results into account as well as the limitations of the analysis methods applied. Based on these results, various directions for future research are provided in Chapter 7.

The Appendix comprises supporting information: Appendix A provides fundamentals for room acoustics, like the RIR or prominent room acoustical parameters. Appendix B addresses some mathematical basics important in spherical sound field descriptions, like special functions, the Fourier transform in spherical coordinates as well as operations for coordinate system rotation. The following Appendices C and D contain additional tables and plots related to the listening tests and prediction analyses.

[2] Note that in acoustics, the terms *subjective* and *objective* are common to distinguish between perceptual and technical aspects, respectively. In [291] it was stressed that subjective listening experiments with subsequent statistical evaluation also provide an objective result. Accordingly, the terms *perceptual* and *technical* (or *physical*) are used throughout this thesis to avoid ambiguity.

2 State of Research: Audio Quality & Quality Prediction

This chapter gives an overview of the state of research in audio quality assessment and prediction. In particular, the acoustic quality of concert halls is addressed, because, in the context of this thesis, microphone arrays are employed for room simulation applications. Therefore, the state of research on quality assessment of VAEs is presented with a focus on VAEs based on microphone array data. The last section of this chapter addresses quality prediction methods commonly applied in concert hall acoustics, audio coding, and spatial audio.

2.1 Quality of Concert Halls

A concert hall, theater, or lecture room may have *good* or *poor* acoustics—populary speaking. For example in a poor-sounding lecture room, the words from a lecturer may not be intelligible, whereas a well-sounding concert hall can significantly increase the overall quality of the performance. The respective quality assessment is subject to concert hall acoustics research which, over the last decades, provided a number of methods and approaches to describe the quality of a room's acoustics. Audio quality—in the sense of concert hall acoustics—can be divided into *timbral* and *spatial* quality, whereas timbral artifacts have been found to be the dominant cue [153]. This means that even a sound field with correctly synthesized spatial features is rated poor in quality if timbral artifacts are audible, like artifcats introduced by the recording or reproduction method (such as high-frequency sounds due to spatial aliasing). Nevertheless, directional sound field properties are known to strongly contribute to the overall quality of a concert hall [250].

Early research established quality attributes such as *intimacy* and *presence* [22], or auditory *spatial impression* [247, 12] and *spaciousness* [140, 107]. Both latter sensations are related to the amount of reverberation in the room [35]. They are quantifiable using the Interaural Time-Delay Gap (ITDG) which represents the delay between direct sound and the first room reflections. In [108], the Average Interaural Time Difference (AITD) and the Diffuse Field Transfer Function (DFTF) were identified as good measures for *externalization* and *envelopment*, respectively, both contributing to listener preference [106]. However, it was pointed out that *spaciousness*, which is assumed to be the same sensation as *envelopment*, is strongly dependent on the presented audio signal. Furthermore, spaciousness can be divided into three individual percepts, namely *continuous spatial impression*, *early spatial impression*, and *background spatial impression.* Experiments in [55] point out that the quality of concert halls is a mixture of three quality dimensions: The first, *ambiance*, is related to *spaciousness* while the second, *clarity*, comprises quality factors such as *intelligibility*, *articulation* or *definition.* The third quality dimension is represented by *loudness.* Another often quoted quality metric is the *auditory scene* or *source width* (ASW) [11], describing the perceived spatial extent of a sound source or scene. Until today, it is used for quality evaluations of concert halls.

Although there is still a debate in the community on what attributes are actually useful for room acoustical quality description, and what physical sound field properties contribute to the perception of spaciousness [24], a consensus was reached regarding two attributes [39, 185, 196], namely ASW and *Listener Envelopment* (LEV). Both are described in more detail in Section 3.2.3 as they are used for qualitative analysis of the array configuration in Section 4.3.

2.2 Quality of Virtual Acoustic Environments

The research on quality assessment of VAEs began in the '90s with the introduction of DBS and loudspeaker-based synthesis systems like WFS or HOA. Although not directly related to VAE quality, the Radiocommunication Sector of the International Telecommunication Union (ITU) proposed two measures for multi-channel surround quality, namely the *front image quality* and the *impression of surround quality* [52]. In addition, metrics from concert hall acoustics were also applied for the evaluation of multi-channel surround systems, which were then further adapted to assess the quality of VAEs. Although doubt has been stated on their applicability in [235] or [260], some of these measures are still widely in use to assess the quality of VAEs, like in the assessment of the spatial impression in stereophonic [239] and surround sound systems [23, 120, 100], in the evaluation of binaural synthesis [15], HOA [31], or WFS [195]. Comprehensive overviews of 3D and multi-channel spatial quality evaluations can for example be found in [236, 166, 235, 144, 255].

In the experiments throughout this thesis, all VAEs are synthesized using binaural technology. Early quality evaluations of binaural auralizations mainly assessed basic auditive criteria, like localization performance or front-back-confusions [180, 178, 159]. Further experiments included integrative quality factors, like *immersion* or *sense of presence* [161, 248], to describe the quality of the reproduction. Although not all attributes were exclusively designed for the evaluation of VAEs, a number of vocabularies were developed and introduced for quality assessment [204, 89, 162, 25, 256]. Recently, Lindau et al. presented the Spatial Audio Quality Inventory (SAQI) to rate the reproduction quality of VAEs [157]. SAQI was developed by means of a focus group approach comprising 13 audio experts and acousticians. It provides various quality attributes to describe different aspects of the synthesis, including temporal, spectral, and spatial characteristics of the presented acoustic scene as well as attributes relating to the technical properties of the reproduction system. With regard to the presented visual environment, findings in [48] stressed that the acoustic reproduction quality in VR applications should be authentic and plausible. Consequently, two attributes, namely *authenticity* and *plausibility*, were proposed for evaluation, whereas the first is related to an externally provided reference and the latter to an imagined (or inner) reference. Recent research investigates scene complexity and room acoustic disparity with their influence on *spatial presence*, *externalization*, *localization accuracy*, and *plausibility* [295].

In the scope of this thesis, the VAEs are based on spherical microphone array recordings. The next section presents an overview of the state of research in quality assessment of (spherical) microphone arrays, when using such systems for auralization applications.

2.3 Quality of Microphone Array Auralizations

In practice, when the auralization is based on microphone array data, then measurement errors such as spatial aliasing or microphone noise significantly degrade the quality of reproduction. Although the influence of the array and system errors have been analyzed and described well from an analytical point of view, at least for spherical microphone arrays [215, 222], only a few recent publications address their perceptual effect.

A perceptual experiment in [168] investigated the influence of measurement errors in circular microphone array auralizations. Rendering the recorded sound field for different stereophonic reproduction, the quality of virtual, array-based microphone setups were compared to real stereophonic setups in a quantitative analysis. A similar experiment was carried out in [167], using spherical arrays for sound field sampling. The perceptual analysis in [203] evaluated four different array configurations for various reproduction setups, using descriptive attributes such as *naturalness*, *envelopment*, *localization*, and *depth*. Based on a FCP approach, experiments in [8] assessed the binaural reproduction quality of spherical microphone array data using various attributes including *timbre balance*, *localization performance* or *transient reproduction accuracy*. The focus of the experiment was on the influence of different spatial accuracies in terms of spherical harmonic orders and the effect of sampling, i.e., spatial aliasing. Investigations in [254] assessed the quality of binaurally auralized spherical array data when applying a timbre correction that significantly increased the reproduction quality. Further experiments in [253] employed generalized spherical array beamforming for binaural speech reproduction, providing better intelligibility compared to other beamforming algorithms. The experiment presented in [193] dealt with the assessment of real spherical array recordings from a concert hall. Using OPQ for attribution, the aim of the analysis was to establish a relation between specific quality factors to overall preference. Specifically, the influence of the array configuration was assessed in a descriptive analysis with naïve listeners. In [28], Bernschütz investigated a variety of relevant aspects in binaural auralizations based on spherical array data. In quantitative analyses and descriptive assessments using SAQI, he provided a comprehensive overview on the influence of various array configurations, the chosen sampling strategy and the impact of error reduction methods on the reproduction quality. The trade-off between generalized spherical array beamforming and the binaural synthesis was investigated in [127], which was based on theoretical simulations and a quantitative perceptual analysis.

2.4 Quality Prediction

As described in Section 1.1, the third component in the proposed approach for quality assessment is predictive modeling which, in general, is a technique to estimate an outcome [141]. It is applied in a variety of disciplines such as social and computer sciences, chemistry, physics, and statistics. To name only a few, some common predictive models are, for example, ordinary models for linear regression, nonlinear models, like artificial neural networks or support vector machines as well as decision trees. Such models relate the perceptual data from listening experiments and the physical sound field measures, hence enabling quality prediction. The following sections present approaches commonly used in concert hall acoustics, audio coding, and spatial audio research.

2.4.1 Quality Prediction in Audio Coding

In early audio coding research, a variety of objective methods were developed to assess the quality of an audio coder at various bit rates. Based on the output variables of an auditory model (MOV), measuring, for example, the amount of coding distortions in the signal, it was possible to relate the perceptual data derived in listening tests, i.e., the quality ratings from assessors, to the MOVs representing the technical parameters of the coded signal. Modeling the relation between technical and perceptual data enabled the prediction of the coding quality without conducting time-consuming listening experiments.

Early methods typically evaluated the quality of speech coders based on SNRs [246, 132] until the segmental Noise-to-Mask-Ratio (NMR) was introduced in [40, 43]. NMR is a measure to evaluate the perceptual amount of coding noise in broadband audio signals, taking the masking properties of the human auditory system into account. Specifically, NMR measures the distance between coding noise and the masking threshold, comparing the coded signal under test to a reference signal. Over the years, further research resulted in a number of technical quality measures employing improved models of internal auditory representations, like in [288, 289]. Prominent quality assessment methods are, for example, the Perceptual Audio Quality Measure (PAQM) [16], the Peripheral Internal Representation (PIR) [272], the Perceptual Evaluation of the Quality of Audio Signals (PERCEVAL) [202], the Perceptual Objective Measure (POM) [62], the Distortion Index (DIX) [275], or the Objective Audio Signal Evaluation (OASE), which provided an improved temporal and spectral resolution [263]. Some of the above mentioned coding quality assessment methods were included in the Perceptual Evaluation of Audio Quality (PEAQ) [276], being standardized as ITU-R BS.1387 [226]. Recently, attempts were made to extend PEAQ towards multi-channel applications as PEAQ-MC [155]. In addition to the original version, PEAQ-MC also takes spatial perception into account (in terms of interaural cues). A further approach for spatial and timbral audio coding evaluation was given in [252]. In ITU-T P.862 [200], a method for the Perceptual Evaluation of Speech Quality (PESQ) [17] in narrowband telephone networks and speech codecs was released [200] which was later extended for wideband applications [201]. In [231], an auditory model was presented for quality prediction of audio coders, introducing the simulation of inner hair cells based on an adaptation circuit. Moreover, quality prediction was applied in wideband speech codecs [213] and voice-over-IP systems [211]. Comprehensive overviews can be found in, e.g., [14, 230, 305].

It should be noted that all these methods perform well solving the specific problem they are designed to solve. However, they do not generalize for other applications.

2.4.2 Quality Prediction in Concert Hall Acoustics

Similar to the prediction of the coding quality, room acoustic quality can also be estimated by connecting perceptual aspects of a sound field to its physical properties. Again, the quality is assessed in perceptual experiments and subsequently related to physical measures taken, for example, from a measured RIR.

Early research mainly related perceptual preference to the direct-to-reverberation ratio to establish physical measures for quality [238]. For example, clarity, C50 or C80, allowed statements on the applicability of a concert hall for speech or music performances, respectively [1]. Further research showed that especially the spatial properties of a sound field strongly contribute to room acoustical quality [250]. It was found that the energy relationship between early and late sound events are unexpectedly critical [105] and that lateral reflections significantly increase spatial quality [107]. Predictors like Lateral Early Decay Time (LEDT), Lateral Energy Fraction (LF), and Late Lateral Energy Level (GLL) correlated well to perception in terms of envelopment [9]. Alternative measures were introduced in [198], namely the Lateral Component (LC), Front-to-Back Ratio (FBR), and the Left-to-Right Ratio (LRR). Further research resulted in a number of spatial quality predictors, like the Bass Index (BI), the Binaural Quality Index (BQI) [133] or the Degree of Source Broadening (DSB) [21], whereas other studies established the sound strength G as a predictor for subjective loudness, which is a useful measure for BI at low frequencies [69, 10, 261]. In [124], G is investigated for concert hall acoustics. The role between perceptual and technical measures is described and established for the early sound field. Experiments in [21] highlighted G, together with the Reverberation Time (RT), as an underestimated quality measure in concert hall acoustics. It was shown that G strongly relates to ASW and LEV. In [303] and [121], LF, G_{mid} (strength at mid-frequencies), and BQI were applied for spatial quality prediction.

Two perceptual quality features, the *frontal spatial fidelity* and the *surround spatial fidelity* have been found to be important measures to describe the quality of concert halls using G, RT, and loudness as predictors [237]. Following the same principle, a prediction model for perceived spaciousness was proposed in [240], taking three spaciousness-dimensions into account, namely *ensemble width*, *extent of reverberation*, and *extent of immersion*. Recent research provided models for the prediction of quality features such as ASW and LEV [78, 137]. Both are based on the evaluation of interaural differences in the early and late part of the sound field, respectively. A comprehensive overview on quality assessment and prediction in concert hall acoustics was recently provided by Weinzierl and Vorländer in [292].

2.4.3 Quality Prediction in Spatial Audio

Quality prediction was also applied in order to assess the quality of multi-channel audio systems or VAEs. This section provides a brief overview while further information on quality assessment and prediction in spatial audio systems are for example given in [255, 191], addressing WFS, HOA, and binaural reproduction.

Experiments in [208] employed an auditory model to estimate the reproduction quality in HOA when changing the amount of loudspeakers used. Based on simple linear regression, a method was presented in [99] to predict the spatial fidelity of a 5.1 system, using G, RT, and loudness as predictors. Also addressing the spatial quality of a 5.1 system, a prediction approach in [87] uses the auditory model developed in [75] to derive predictors which are subsequently related to the results from a listening test using multivariate adaptive regression splines. Based on 5-channel audio recordings derived in [64], an regression-based approach was presented in [65] to predict the quality of 5.1 systems in terms of source location, envelopment or spaciousness, and timbre. Specifically, the Quality Evaluation of Spatial Transmission and Reproduction using an Artificial Listener (QESTRAL) was developed in [63], using partial least squares regression and neural networks for prediction. Experiments in [58] and [57] estimated the spatial fidelity of various reproduction systems such as stereo, Dolby Pro Logic II, DTS Neo:6, and 5.0, using simple regression to relate perceptual and technical data. In [210], the quality of stereo, 5.1, and HOA systems as well as binaural technology was estimated based on the binaural signals of a dummy head in order to assess different headphone types for spatial audio playback. The recently published PhD thesis by Wierstorf deals with the prediction of the spatial reproduction quality in HOA and WFS systems [297]. Specifically, it estimates the influence of the employed loudspeaker setup on spatial fidelity and the size of the sweet spot (also see [273]).

Quality predictions of binaural auralizations are commonly based on evaluating the differences in the binaural signals, like in [257]. The experiments addressed the interaction of the listener in binaural auralizations, evaluating different HRTF data bases. In [209], different panning techniques were assessed using a binaural auditory model to evaluate localization cues and colorations.

Although in [233], a quality prediction approach was presented using a microphone array in binaural hearing aids, so far only one publication by the author of this thesis addresses the binaural quality and its prediction in spherical microphone array auralizations: In [194], the auditory model developed in [137] was applied to estimate the reproduction quality in terms of ASW and LEV.

3 Theoretical Background

This chapter provides an overview of the theoretical background concerning the three basic fields of research underlying the studies presented in this thesis. Section 3.1 presents the fundamentals of spherical sound fields including special functions, sampling strategies, microphone array configurations, and system errors. Section 3.2 addresses psychoacoustics, particularly the human auditory system, binaural hearing, the perception of spaciousness, and the auditory model used in PEAQ. Section 3.3 deals with predictive modeling, describing common models and methods for model tuning and evaluation as well as prominent problems.

3.1 Spherical Sound Fields

This section presents the fundamentals for sound field descriptions in spherical coordinates, mainly following [218]. Such sound fields are commonly represented using spherical basis functions, the so-called spherical harmonics. They are introduced first as a fundamental tool in spherical array processing. The subsequent sections describe sound field sampling of order-limited functions on a sphere, i.e., sampling sound pressures with a spherical microphone array, discussing common sampling schemes, array configurations, and their properties. Then, spatial filtering, i.e., beamforming in spherical microphone arrays, is introduced with a focus on the PWD which is the beamforming approach used in the experiments in this thesis. Furthermore, measurement errors are analytically described which degrade the spatial response of the array in practice, specifically addressing spatial aliasing, measurement noise, and microphone positioning errors. The section closes with a brief description of HRTFs represented in spherical coordinates.

Sound field formulations representing an acoustic wave field are generally based on solutions of the acoustical wave equation. The spherical wave equation, i.e., the Helmholtz equation, and its derivation are given in Appendix B.1. Its solutions are represented by spherical basis functions such as spherical harmonics as angular solutions as well as spherical Hankel and Bessel functions as radial solutions (see Appendix B.2).

3.1.1 Spherical Harmonics

Fundamental to spherical sound field descriptions are the spherical harmonics Y_n^m of order $n \in \mathbb{N}$ and mode $m \in \mathbb{Z}$. Accordingly, a sound field is represented by a weighted sum of a set of spherical harmonics which are defined as [299]

$$Y_n^m(\theta, \phi) \equiv \sqrt{\frac{2n+1}{4\pi}\frac{(n-m)!}{(n+m)!}}P_n^m(\cos\theta)e^{im\phi}, \tag{3.1}$$

with i being the imaginary unit, $(\cdot)!$ the factorial function, and $P_n^m(\cdot)$ the associated Legendre functions (see Appendix B.2). The angles in elevation and azimuth direction are denoted θ and ϕ, respectively. The order n controls the dependence of the spherical harmonics over θ via $\sin\theta$ and $\cos\theta$ while the mode m relates to ϕ through $e^{im\phi}$ [218]. Figure 3.1 shows a set of spherical harmonics for orders $n = 0 \ldots 2$ and modes $m = -2 \ldots 2$.

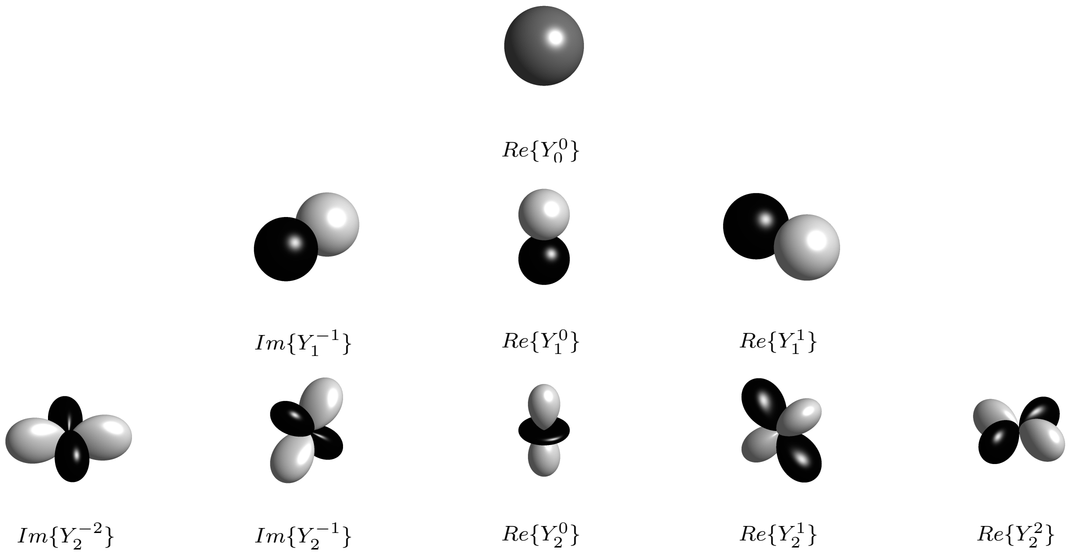


Figure 3.1: Set of spherical harmonics of orders $n = 0 \ldots 2$. The gray areas mark the negative values, the black, the respective positive values.

The two different color values, i.e., black and gray, indicate negative and positive values, respectively. Note the directional behavior of a single spherical harmonic function. For example, order $n = 0$ shows a monopole behavior while order $n = 1$ provides a dipole characteristic. The higher the sound field order is calculated, the more modes and side-lobes arise, offering an improved directional response for higher frequencies.

The spherical harmonics form the Fourier basis for functions on the sphere, providing an intuitive tool for spherical array processing. For example, a function $f(\theta, \phi) \in L_2(S^2)$ can be represented with a weighted sum of spherical harmonics, which can be formulated by [218]

$$f(\theta, \phi) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} f_{nm} Y_n^m(\theta, \phi). \tag{3.2}$$

$L_2(S^2)$ is the Hilbert space comprising a set of all square-integrable functions on the unit sphere such as the spherical harmonics Y_n^m. f_{nm} are the weights representing the

$$f_{nm} = \int_0^{2\pi} \int_0^{\pi} f(\theta, \phi)[Y_n^m(\theta, \phi)]^* \sin\theta d\theta d\phi, \tag{3.3}$$

with the asterisk $*$ denoting the conjugate complex. The infinite spherical harmonics series represents the Dirac delta distribution over the sphere around (θ', ϕ') with Fourier coefficients $\left[Y_n^m(\theta', \phi')\right]^*$.

If the series is truncated to a finite order N, then a closed form expression can be written as, after [220]:

$$\begin{aligned} f(\theta, \phi) &= \sum_{n=0}^{N} \sum_{m=-n}^{n} \left[Y_n^m(\theta', \phi')\right]^* Y_n^m(\theta, \phi) \\ &= \frac{N+1}{4\pi(\cos\Theta - 1)} \left[P_{N+1}(\cos\Theta) - P_N(\cos\Theta)\right], \end{aligned} \tag{3.4}$$

with Θ being the angle between θ and θ', and P_N representing the Legendre polynomials. This directional behavior of a truncated spherical harmonics is exemplarily shown in Figure 3.2 a) for orders $N = 1, 10$ and 40, clearly illustrating the improved directional response for higher orders.

The relation between the PWD and the employed sound field order N is depicted in Figure 3.2 b), considering the first (smallest) zero of Θ_0 of $w_N(\Theta)$, after [220]. The plot shows Θ_0 as a function of N (solid line), together with its approximation (dashed line) defined by [220]

$$\Theta_0 \approx \frac{\pi}{N}. \tag{3.5}$$

The approximation error is less than 2^o for $N \in [4, 40]$. The resolution of the PWD is a common measure of the array performance. Based on spherical harmonics, the representation of fundamental wave types in spherical coordinates, like plane and point source sound fields, is described next.

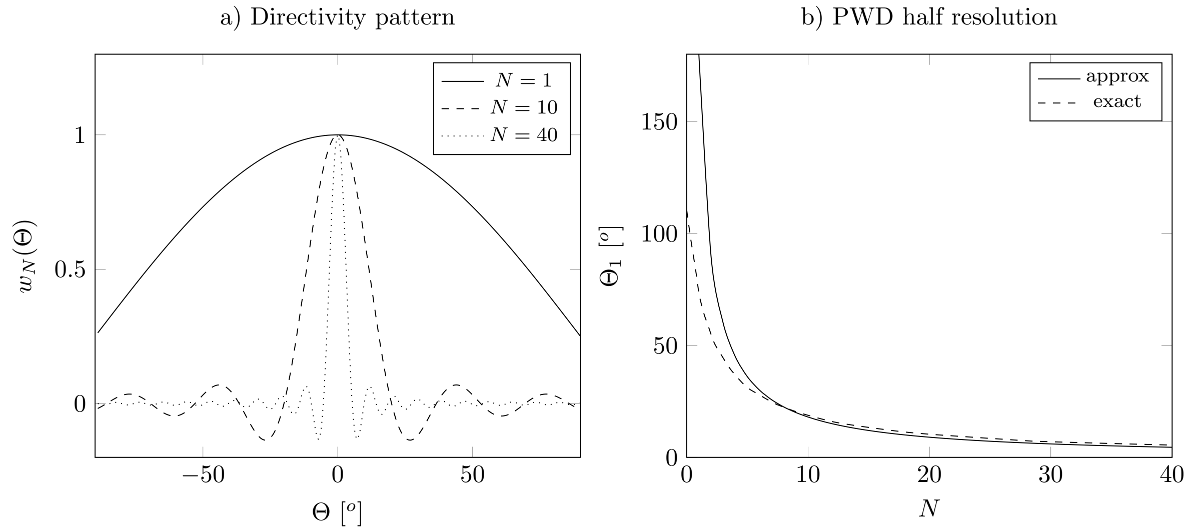


Figure 3.2: Directivity pattern of truncated spherical harmonics series for orders $N = 1, 10$ and 40 in plot a) and the exact and approximated half resolution of the PWD in plot b) as solid and dashed lines, respectively.

3.1.2 Fundamental Waves in Spherical Coordinates

Plane waves and sound fields originating from point sources, i.e., spherical waves, are fundamental wave types commonly used in acoustic sound field descriptions. In Cartesian coordinates, they both are solutions to the wave equation, hence, representing a sound field. The following sections address the spherical representations of plane and spherical wave fields which provide a solution to the Helmholtz equation, i.e., the wave equation in spherical coordinates. The descriptions follow [218].

Plane waves

A single-frequency plane wave originating in direction (θ_k, ϕ_k) is represented by the wave vector $\tilde{\boldsymbol{k}} = -\boldsymbol{k} = (k, \theta_k, \phi_k)$. The wavenumber k is related to frequency by $k = 2\pi f/c$, with c denoting the speed of sound in the respective medium. The sound pressure p at $\boldsymbol{r} = (r, \theta, \phi)$ can be written as a summation of spherical harmonics and Bessel functions, as described by

$$\begin{aligned} p(k, r, \theta, \phi) &= e^{-i\boldsymbol{k}\cdot\boldsymbol{r}} = e^{i\tilde{\boldsymbol{k}}\cdot\mathbf{r}} \\ &= \sum_{n=0}^{\infty}\sum_{m=-n}^{n} 4\pi i^n j_n(kr)[Y_n^m(\theta_k, \phi_k)]^* Y_n^m(\theta, \phi) \end{aligned} \tag{3.6}$$

with i being the imaginary unit and $(\cdot)$ the dot product which is given by $\tilde{\boldsymbol{k}} \cdot \boldsymbol{r} = kr\cos\Theta$, with Θ representing the angle between θ and θ'. Note that the advantage of a plane wave representation in the spherical harmonics domain is the possibility to separate variables into kr, wave arrival direction (θ_k, ϕ_k), and position on the surface of the sphere (θ, ϕ). This provides a flexible basis for various array processing algorithms which are presented later in this section, following [218].

If the sound field of a single plane wave is described on the surface of a sphere with radius r, then the function on the sphere, i.e., the measured sound pressure $p(k, r, \theta, \phi)$, satisfies

$$p(k, r, \theta, \phi) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} p_{nm}(k, r) Y_n^m(\theta, \phi), \tag{3.7}$$

with $p_{nm}(k, r)$ representing the spherical Fourier coefficients. The respective coefficients for a unit-amplitude plane wave arriving from (θ_k, ϕ_k) can then accordingly be formulated by

$$p_{nm}(k, r) = 4\pi i j_n(kr)[Y_n^m(\theta_k, \phi_k)]^*. \tag{3.8}$$

Note that the magnitude of p_{nm} is proportional to the magnitude of $j_n(kr)$, indicating that a plane wave sound field decays as a function of n for $n \geq kr$ (see [218] for details). The Fourier coefficients p_{nm} are also referred to as the *spherical wave spectrum*, like for example described in [299].

Plane wave composition: If a sound field is composed of multiple or an infinite number of plane waves, then it can be represented as a summation over the plane wave term in Eq. (3.6), resulting in

$$p(k, r, \theta, \phi) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} 4\pi i^n a_{nm}(k) j_n(kr) Y_n^m(\theta, \phi), \tag{3.9}$$

with $a_{nm}(k)$ being the spherical Fourier transform of the directional amplitude density $a(k, \theta_k, \phi_k)$. With respect to Eqs. (3.6) and (3.9), the expression $a_{nm}(k) = [Y_n^m(\theta_k, \phi_k)]^*$ holds and consequently, following Eq. (3.9), the plane wave sound field at the surface of a sphere with radius r can be expressed in the spherical harmonics domain by

$$p_{nm}(k, r) = 4\pi i^n a_{nm}(k) j_n(kr). \tag{3.10}$$

However, the infinite summation in Eq. (3.9) can in practice be approximated by a finite summation, replacing the ∞ symbol with N. Based on Eqs. (3.9) and (3.8), the sound pressure at an arbitrary point in space $(r^{'}, \theta^{'}, \phi^{'})$ can be calculated, extracting $a_{nm}(k)$ through division by $4\pi i^n j_n(kr)$, and reconstructing $p_{nm}(k, r)$ by multiplication with $4\pi i^n j_n(kr^{'})$, which results in

$$p(k, r^{'}, \theta^{'}, \phi^{'}) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \frac{j_n(kr^{'})}{j_n(kr)} p_{nm}(k, r) Y_n^m(\theta^{'}, \phi^{'}). \tag{3.11}$$

If the infinite sum is approximated by a finite order summation, then the order-limited equation is only useful in a range, where kr and $kr^{'}$ are smaller than N. Following Eq. (3.11), the pressure field is given by

$$p(k, r^{'}, \theta^{'}, \phi^{'}) \approx \sum_{n=0}^{N} \sum_{m=-n}^{n} \frac{j_n(kr^{'})}{j_n(kr)} p_{nm}(k, r) Y_n^m(\theta^{'}, \phi^{'}). \tag{3.12}$$

Note the Bessel function $j_n(kr)$ in the denominator of Eqs. (3.11) and (3.12). In practice, numerical instabilities arise due to a division by zero, when kr values correspond to the zeros of the Bessel functions, or when their magnitude becomes low for $n > kr$.

Spherical waves

Plane wave sound fields are quite rare in real life acoustics. A realistic sound source is commonly modeled as a point source, i.e., a monopole source, which produces a spherical pressure field with a magnitude decaying inversely proportional to the distance of the source. The phase, however, describes a constant function over θ and ϕ. For a sound source located at $\boldsymbol{r}_s = (r_s, \theta_s, \phi_s)$, the pressure at location $\boldsymbol{r} = (r, \theta, \phi)$ can be calculated by means of a spherical harmonics series [299]

$$\frac{e^{-ik\|\boldsymbol{r}-\boldsymbol{r}_s\|}}{\|\boldsymbol{r}-\boldsymbol{r}_s\|} = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} 4\pi(-i)k h_n^{(2)}(kr_s) j_n(kr) [Y_n^m(\theta_s, \phi_s)]^* Y_n^m(\theta, \phi), \quad r < r_s, \tag{3.13}$$

with $\|\boldsymbol{r}\| = r$ and $\| \cdot \|$ being the Euclidean norm. $h_n^{[2]}$ is the Hankel function of the second kind, which is described in Appendix B.2. In this example, the point source is located outside of the measurement sphere with radius r. This is the so-called *interior problem* which is described in detail in [299]. Similarly, if condition $r > r_s$ is fulfilled, then the sound source is within the measurement sphere, i.e., an *exterior problem*, which is expressed by

$$\frac{e^{-ik\|\boldsymbol{r}-\boldsymbol{r}_s\|}}{\|\boldsymbol{r}-\boldsymbol{r}_s\|} = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} 4\pi(-i)k h_n^{(2)}(kr) j_n(kr_s) [Y_n^m(\theta_s, \phi_s)]^* Y_n^m(\theta, \phi), \quad r > r_s. \tag{3.14}$$

Such an approach is used, for example, when measuring the directivity characteristics of a sound source, as described in [308]. However, in this thesis, the scope of investigation is on microphone arrays used for room recordings, hence, addressing only the interior problem. For further information on exterior problems, refer to [299, 308].

For the interior problem described above, i.e., with $r < r_s$, the pressure on the sphere $p(r, k, \theta, \phi)$ can be represented in the spherical harmonics domain relating Eqs. (3.13) and (3.7), which leads to

$$p_{nm}(k, r) = 4\pi(-i)kh_n^{(2)}(kr_s)j_n(kr)[Y_n^m(\theta_s, \phi_s)]^*, \quad r < r_s. \tag{3.15}$$

So far, plane and spherical wave sound fields represented by spherical harmonics were reviewed. In the following, the respective sound field sampling is described based on spherical microphone arrays. In this regard, various sampling schemes and common array configurations are presented.

3.1.3 Sampling Order-Limited Functions

When a continuous function is sampled and its perfect reconstruction is desired, then the sampling theorem, also known as the Nyquist theorem [130], has to be fulfilled. This can be achieved, for example, by sampling band-limited or, as for sound fields represented by spherical harmonics, order-limited functions. Based on so-called *quadrature* methods, the integral of a given function $g(\theta, \phi)$ can be approximated by a sum over all samples (θ_q, ϕ_q) of this function, as described by

$$\int_0^{2\pi}\int_0^{\pi} g(\theta, \phi)\sin\theta d\theta\ d\phi \approx \sum_{q=1}^{Q} \alpha_q g(\theta_q, \phi_q). \tag{3.16}$$

The sampling weights are denoted α_q, and Q is the total number of samples. After [218], the SFT of function $f(\theta, \phi)$ can be approximated, extending the formulation by substituting $g(\theta, \phi) = f(\theta, \phi)[Y_n^m(\theta, \phi)]^*$, which leads to

$$\begin{aligned} f_{nm} &= \int_0^{2\pi}\int_0^{\pi} f(\theta, \phi)[Y_n^m(\theta, \phi)]^* \sin\theta d\theta d\phi \\ &\approx \sum_{q=1}^{Q} \alpha_q f(\theta_q, \phi_q)[Y_n^m(\theta_q, \phi_q)]^*. \end{aligned} \tag{3.17}$$

If Q is sufficiently high, then $f(\theta, \phi)$ can be perfectly reconstructed using the inverse SFT (ISFT). The SFT and spherical convolution are described in Appendix B.3 and B.4, respectively. Based on the orthogonality of the spherical harmonics (see Eq. (B.32)), and substituting $f(\theta, \phi)$ with $Y_{n'}^{m'}(\theta_q, \phi_q)$, Eq. (3.17) can be reduced to

$$\sum_{q=1}^{Q} \alpha_q Y_{n'}^{m'}(\theta_q, \phi_q)[Y_n^m(\theta_q, \phi_q)]^* \approx \delta_{nn'}\delta_{mm'}, \tag{3.18}$$

with $\delta_{nn'} = 1$ for $n = n'$, and $\delta_{nn'} = 0$ otherwise. Some common sampling strategies in spherical microphone arrays are described next.

Sampling quadratures

In case of order-limited functions, these sampling methods provide closed-form expressions to compute the Fourier transform of a function on the sphere. However, in practical applications, this requirement may not be fulfilled, leading to sampling errors, like spatial aliasing. Figure 3.3 exemplarily shows the three quadratures for a sound field order $N = 8$, facilitating a similar number of microphones distributed on the sphere.

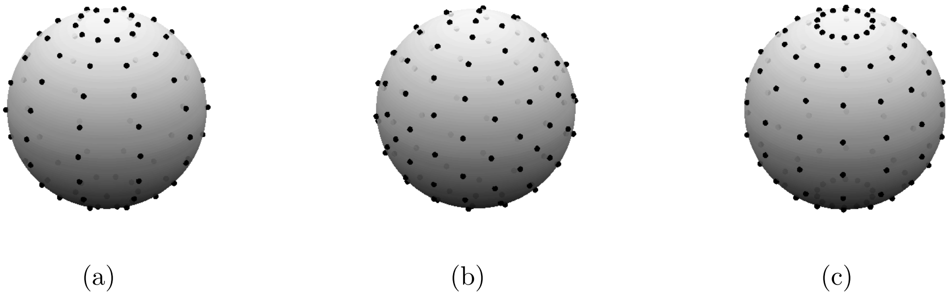


Figure 3.3: Sampling quadratures, showing the equi-angle Chebyshev-grid with 110 samples in a), the Lebedev-grid (uniform) with 110 samples in b), and the Gaussian sampling quadrature with 112 samples in c).

Although all these quadratures provide the same sound field order, their impact on perception can be different in certain cases, as was shown in [28]. In the following, three sampling quadratures commonly used in spherical microphone array sampling are presented, namely equal-angle, Gaussian, and (nearly) uniform sampling, following [218].

Equal-angle sampling: The most intuitive sampling scheme uniformly samples the sphere in both azimuth and elevation direction with $\phi \in [0, 2\pi)$ and $\theta \in [0, \pi]$, respectively. For a given maximum order N, this scheme requires $Q = 4 \cdot (2N + 1)$, after [218]:

$$
\begin{aligned}
\theta_q &= \left(q+\frac{1}{2}\right)\frac{\pi}{2N+2}, & q &= 0,\dots,2N+1 \\
\phi_l &= l\frac{2\pi}{2N+2}, & l &= 0,\dots,2N+1.
\end{aligned}
\tag{3.19}
$$

Based on delta distributions, equal-angle sampling can be formulated by

$$
s(\theta,\phi) = \sum_{q=0}^{2N+1}\sum_{l=0}^{2N+1}\alpha_q\delta(\cos\theta-\cos\theta_q)\delta(\phi-\phi_q). \tag{3.20}
$$

The weights α_q determine the amplitude of the delta distribution which reduces towards the poles due to the increased sample density. After [218], the weights are calculated by

$$
\alpha_q = \frac{2\pi}{(N+1)^2}\sin(\theta_q)\sum_{q^{'}=0}^{N}\frac{1}{2q^{'}+1}\sin([2q^{'}+1]\theta_q), \qquad 0 \le q \le 2N+1. \tag{3.21}
$$

The SFT of $s(\theta,\phi)$, which is derived following [218], can be written as

$$
s_{nm} = \sqrt{4\pi}\delta_n\delta_m + \tilde{s}_{nm}, \tag{3.22}
$$

with $\tilde{s}_{nm}$ is non-zero for $n > 2N+1$. A sampled function on the sphere, as defined by $f_s(\theta,\phi) = f(\theta,\phi)s(\theta,\phi)$, can be written as

$$
f_s(\theta,\phi) = f(\theta,\phi) + f(\theta,\phi)\tilde{s}(\theta,\phi), \tag{3.23}
$$

with $\tilde{s}(\theta,\phi)$ being the ISFT of $\tilde{s}_{nm}$. With $f(\theta,\phi)$ and $\tilde{s}_{nm}(\theta,\phi)$ being order-limited to $n \le N$ and $n \ge 2N+2$, respectively, the SFT of product $f(\theta,\phi)\tilde{s}(\theta,\phi)$ has zero coefficients for $n \le N+1$, leading to $f_{snm} = f_{nm}$. Now function $f(\theta,\phi) = Y_n^m(\theta,\phi)$ can be reconstructed from the sampled function $f_s(\theta,\phi)$ without aliasing, employing a spatial low-pass filter $h(\theta,\phi)$ with cut-off order N. Finally, f_{nm} can be formulated as

$$
f_{nm} = 2\pi\sqrt{\frac{4\pi}{2n+1}}f_{snm}h_{n0} \quad = \begin{cases} f_{snm} & n \le N \\ 0 & \text{otherwise.} \end{cases} \tag{3.24}
$$

The Chebychev quadrature [96, 167], as depicted in Figure 3.3 c), is such an equal-angle sampling rule which, due to its regular lattice, is commonly employed when using scanning microphone arrays for sound field sampling. A disadvantage of this sampling scheme is the relatively high number of samples needed, compared to other quadratures.

Uniform and nearly-uniform sampling: In order to distribute sampling points uniformly on the surface of a sphere, the sphere is separated into polyhedra, the so-called Platonic solids. Their vertices represent the respective sample positions, satisfying the quadrature relation (after [115])

$$\int_0^{2\pi}\int_0^{\pi} g(\theta,\phi)\sin\theta d\theta\ d\phi = \frac{4\pi}{Q}\sum_{q=1}^{Q} g(\theta_q,\phi_q), \tag{3.25}$$

with $\alpha_q = 4\pi/Q$ being the sampling weights with regard to Eq. (3.16). If spherical arrays are employed to sample order-limited sound fields, then, following [218], Eq. (3.25) can be rewritten as

$$f_{nm} = \frac{4\pi}{Q}\sum_{q=1}^{Q} f(\theta_q,\phi_q)Y_n^m[(\theta_q,\phi_q)]^*, \tag{3.26}$$

when $g(\theta,\phi)$ in Eq. (3.25) is replaced by $f(\theta_q,\phi_q)Y_n^m[(\theta_q,\phi_q)]^*$, as formulated in Eq. (3.17). The Lebedev quadrature, for example, is a common sampling scheme which is characterized by samples nearly uniformly distributed on the sphere [150, 148, 149]. Since all samples have the same distance to their nearest neighbors, the lattice is separable neither in azimuth nor in elevation direction. Consequently, no general formula to derive the sample coordinates is given. However, Fortran code to calculate grids for orders up to $N = 131$ have been provided in [147].

When Lebedev sampling is employed, then a sound field representation for a given order N can be achieved with $Q \approx \frac{(N+1)^2}{3}$ samples. Compared to the other quadratures, the Lebedev grid offers the least number of samples for a given order and is therefore the sampling scheme of choice in practical applications. In addition, it also provides an improved robustness against spatial aliasing as will be shown in the simulation examples provided in Section 3.1.5.

Gaussian sampling: The Gaussian sampling scheme is characterized by sampling at the zeros of the Legendre polynomials $P_{N+2}^2(\cos\theta_q) = 0$, using $N+1$ samples in elevation direction. This satisfies the orthogonality constraint over the summation of the Legendre functions

$$\sum_{q=0}^{N} \alpha_q P_n(\cos\theta_q) = \frac{2\pi}{N+1}\delta_n, \qquad n \leq 2N+1. \tag{3.27}$$

The azimuth direction is uniformly sampled with $2(N+1)$ samples. Consequently, a sound field can be sampled up to order N using $2(N+1)^2$ samples. After [218], the spherical Fourier coefficients can be derived by

$$f_{nm} = \sum_{q=0}^{N} \sum_{l=0}^{2N+1} \alpha_q f(\theta_q, \phi_q) Y_n^m \left[(\theta_q, \phi_q)\right]^*, \qquad n \leq N. \tag{3.28}$$

The respective weights α_q are given by

$$\alpha_q = \frac{\pi}{N+1} \frac{2(1-\cos^2\theta_q)}{(N+2)^2 P_{N+2}^2(\cos\theta_q)}, \qquad 0 \leq q \leq N. \tag{3.29}$$

An advantage of the Gaussian grid is the lower amount of samples needed, compared to equal-angle sampling. Further information on sampling strategies is given in [96, 309].

Spherical array configurations

In practical applications, different spherical microphone array configurations are applicable for sound field sampling. Three most commonly employed designs are presented next, employing microphones on a transparent lattice, namely open-omni and open-cardioid sphere configurations comprising omni-directional and cardioid directivity characteristics, respectively. The third, most robust design is the rigid sphere configuration which accounts for sound scattering on the rigid surface of the sphere in its sound field description. In addition, other array configurations are proposed in the literature, like dual-radii designs [169, 20] which sample the sound field on two concentric spheres for bandwidth extension and error robustness improvement, or array configurations which are based on numerical array design, allowing a more flexible microphone placement on a sphere [218].

However, in the following, the three above mentioned array configurations are presented, i.e., the open-omni and open-cardioid designs as well as the rigid sphere configuration. The latter is the configuration used in the experiments throughout this thesis since it provides some advantages compared to other configurations, as will be shown later. Based on simulation examples, all three configurations are evaluated in Section 3.1.5 in terms of their robustness against measurement errors. For more detailed information on spherical array signal processing as well as (numerical) array design and improvement, the reader is referred to [218].

Open-omni sphere: An open sphere array of order N is assumed, employing Q pressure microphones, i.e., microphones with an omni-directional directivity characteristic. The microphones are positioned on the surface of a sphere with radius r, according to the chosen sampling scheme as described in Section 3.1.3. From the sound pressures

$p(k, r, \theta_q, \phi_q)$, measured at each sample position, the respective spherical harmonic coefficients p_{nm} can be calculated by

$$p_{nm}(k, r) = \sum_{q=1}^{Q} \alpha_q^{nm} p(k, r, \theta_q, \phi_q), \qquad n \leq N, \tag{3.30}$$

with α_q^{nm} being the sampling weights and k the wavenumber. If the sound field is order-limited, i.e., $p_{nm} = 0 \ \forall n > N$, then the pressure field on the sphere can be perfectly reconstructed. If the sound field is not order-limited, spatial aliasing occurs for higher order components which, in practice, is unavoidable. However, the effect of spatial aliasing can be reduced, as will be described in Section 3.1.5, which provides detailed information on spatial aliasing and other system errors. The pressure reconstruction for a position outside the measurement sphere $(r^{'}, \theta^{'}, \phi^{'})$, i.e., $r < r^{'}$ is described by

$$p(k, r^{'}, \theta^{'}, \phi^{'}) = \sum_{n=0}^{N} \sum_{m=-n}^{n} \frac{j_n(kr^{'})}{j_n(kr)} p_{nm}(k, r) Y_n^m(\theta^{'}, \phi^{'}). \tag{3.31}$$

From Eq. (3.31), it becomes clear that the reconstruction is only possible if $j_n(kr) \neq 0$. This is the main disadvantage of open-omni sphere designs as a division by zero (or very small values) is difficult to avoid in practice. As illustrated by the zeros of the Bessel functions in Plot a) of Figure 3.4, they are represented by the notches at high kr.

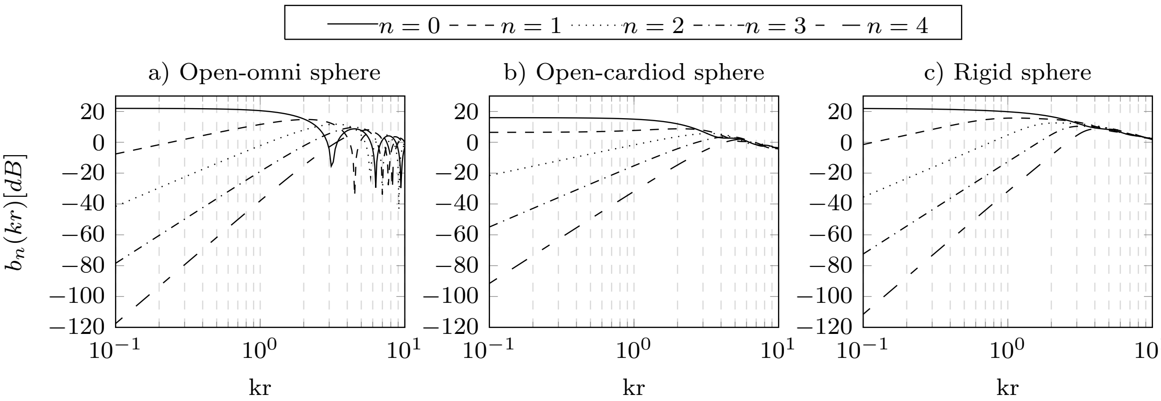


Figure 3.4: Mode strength $b_n(kr)$ in dB as a function of kr for orders $n = 0 \ldots 4$ for open sphere arrays with omni and cardioid microphones (depicted in plots a) and b), respectively). Plot c) shows $b_n(kr)$ for a rigid sphere array.

The division by zero is also an issue when measurement noise is present, like uncorrelated noise from the microphones, which may significantly amplify low frequencies when reconstructing the sound field for higher n at low kr. It is discussed in more detail in Section 3.1.5.

In order to evaluate the performance of a given array configuration, it is useful to relate the sound field, i.e., the plane wave a_{nm}, to the measurement, i.e., the pressure p_{nm} on the sphere, as in Eq. (3.10):

$$p_{nm}(k, r) = a_{nm}(k) b_n(kr). \tag{3.32}$$

$b_n(kr)$ is the mode strength which for open-omni sphere configurations reads

$$b_n(kr) = 4\pi i^n j_n(kr). \tag{3.33}$$

It is shown in Figure 3.4 a) for orders $n = 0 \dots 4$. Note the low pass characteristic for order $n = 0$, and the band-pass behavior with increasing slope for higher orders. At small kr, i.e., at low frequencies, the zero level is present and levels satisfying $n < kr$ should have sufficient power. Levels for $n > kr$, on the contrary, rapidly loose amplitude. Therefore, it is recommended to operate the array at a maximum order $N_{max} = kr$. In practical applications, the operational bandwidth of the array is limited at low frequencies by numerical ill-conditionings due to low values of b_n, and at high frequencies by spatial aliasing. In [223, 169, 20], approaches for bandwidth extensions were proposed, based on dual-radii designs. Here, the idea is to retrieve missing information of one array from the other by cross-fading between the spectra. For example, considering a dual-radius design with $r_1 > r_2$, the corrupted low frequency range of the array with radius r_2 can be reconstructed from the array with radius r_1.

Open-cardioid sphere: The open-omni sphere design presented earlier lacks in robustness due to the nulls of the Bessel functions. A way to compensate for this effect is to use microphones with a cardioid directivity characteristic, hence the term open-cardioid sphere. Such array configurations were subject to investigations in recent publications, like for example in [169, 222].

When using an open-cardioid sphere array for sound field sampling, then the sound pressure measured with a cardioid microphone can be written as

$$x(k, r, \theta, \phi) = p(k, r, \theta, \phi) + \frac{1}{ik}\frac{\delta}{\delta r} p(k, r, \theta, \phi). \tag{3.34}$$

Accordingly, the response to a unit-amplitude plane wave can be formulated by

$$x(k, r, \theta, \phi) = e^{ikr\cos\Theta}(1 + \cos\Theta), \tag{3.35}$$

substituting $p(k,r,\theta,\phi) = e^{i\boldsymbol{k}\cdot\boldsymbol{r}} = e^{ikr\cos\Theta}$ in Eq. (3.34), with Θ denoting the angle between the plane wave direction of arrival and the array look direction. In the spherical harmonics domain, Eq. (3.34) can be rewritten as

$$x_{nm}(k,r) = 4\pi i^n \left[j_n(kr) - ij_n^{'})(kr) \right] \left[Y_n^m(\theta_k,\phi_k) \right]^* . \tag{3.36}$$

Likewise, a sound field comprising a continuum of plane waves with amplitudes $a(k,\theta,\phi)$ reads $x_{nm}(k,r) = b_n(kr)a_{nm}(k)$, with

$$b_n(kr) = 4\pi i^n \left[j_n(kr) - ij_n^{'}(kr) \right] \tag{3.37}$$

being the mode strength for open-cardioid sphere arrays. It is depicted in Plot b) of Figure 3.4 for orders $n = 0\ldots4$.

Rigid sphere: In practical applications, the rigid sphere configuration is commonly used, distributing the microphones on a rigid, i.e., a fully reflecting surface [170]. As shown in [215], the rigid sphere array offers the highest robustness against spatial aliasing and requires the lowest number of microphones for a given order N, compared to other configurations. The sound pressure around a rigid sphere is composed of the free-field incident sound field p_i and the sound field scattered from the rigid surface of the sphere p_s. With a sphere of radius r_a, a boundary condition on its surface is imposed of zero radial velocity due to the infinite impedance at the sphere's boundary. After [218], it can be written as

$$u_r(k,r_a,\theta,\phi) = 0. \tag{3.38}$$

Acoustic velocity is related to pressure through the Euler equation, i.e., the equation of momentum conservation, which reads in spherical coordinates

$$i\rho_0 cku(k,r,\theta,\phi) = \nabla p(k,r,\theta,\phi). \tag{3.39}$$

The gradient operator is defined as

$$\nabla p \equiv \frac{\delta p}{\delta r}\hat{r} + \frac{1}{r}\frac{\delta p}{\delta\theta}\hat{\theta} + \frac{1}{r\sin\theta}\frac{\delta p}{\delta\phi}\hat{\phi}, \tag{3.40}$$

ρ_0 is the air density, and $\hat{r}$, $\hat{\theta}$, $\hat{\phi}$ are unit vectors. Substituting Eqs. (3.38) and (3.40) in Eq. (3.39), the pressures for the scattered and incident sound field can be derived, providing the total pressure formula around the rigid sphere

$$p(k, r, \theta, \phi) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} a_{nm}(k) 4\pi i^n \left[j_n(kr) - \frac{j_n^{'}(kr_a)}{h_n^{(2)'}(kr_a)} h_n^{(2)}(kr) \right] Y_n^m(\theta, \phi). \tag{3.41}$$

Consequently, as in Eq. (3.32), the mode strength of a rigid sphere array configuration can be written as

$$b_n(kr) = 4\pi i^n \left[j_n(kr) - \frac{j_n^{'}(kr_a)}{h_n^{(2)'}(kr_a)} h_n^{(2)}(kr) \right], \tag{3.42}$$

with r_a being the sphere's radius and r denoting the distance to a point on or outside the rigid sphere surface, satisfying $r \geq r_a$. Plot c) of Figure 3.4 shows b_n for a rigid sphere array for orders $n = 0 \ldots 4$, with $r = r_a$. Note that no zeros can be found away from the origin, which is an important property when applying array processing.

3.1.4 Spherical Array Beamforming

Once a spatial sound field is sampled with a spherical microphone array, array processing algorithms can be applied to the input signals. This results in a single output signal which can be designed to have any desired directivity characteristic, ranging between an omni-directional and a figure-of-eight shape. A common processing approach is directional filtering, i.e., beamforming, which allows to focus the sensitivity of the array to any direction of interest, hence, suppressing sound from unwanted directions. A beamforming spherical array was, for example, used in binaural speech reproduction in [253]. Various approaches for beam pattern design exist in the literature which are briefly presented in the following. However, the focus is on the PWD as it is the beamforming approach used in the experiments throughout this thesis. In this regard, the directivity index (DI) and the white noise gain (WNG) are introduced as beamforming performance measures. For more detailed information on optimal beam pattern design and noise minimization, the reader is referred to [44, 279, 218].

Array performance measures

In order to quantify the performance of a beamforming array, two measures are commonly used: DI and WNG. The first measures the directional performance while the second is a robustness measure of the array against measurement noise and uncertainties in system parameters.

Directivity index: The DI (in dB) quantifies the ratio between peak and average values of the squared beam pattern. It can be interpreted as the SNR improvement due to the directional response of the array. DI is computed by $\mathrm{DI} = 10\log_{10}(\mathrm{DF})$, where DF is the directivity factor which is defined as

$$\mathrm{DF} = \frac{|y(\theta_l, \phi_l)|}{\frac{1}{4\pi}\int_0^{2\pi}\int_0^{\pi} |y(\theta, \phi)|^2 \sin\theta d\theta d\phi}. \tag{3.43}$$

White noise gain: To provide a formulation for WNG, a beamforming array looking towards the direction of arrival of a unit-amplitude plane wave is assumed. Uncorrelated noise with variance of σ^2 and zero mean is added to the microphone signals for WNG expression. With the array output in response to the plane wave being $|y|^2 = |\boldsymbol{w_{nm}}^H \boldsymbol{v_{nm}}|^2$, the noise variance at the array output is described by

$$E[|y|^2] = E[yy^H] = \boldsymbol{w_{nm}}^H E[\boldsymbol{p_{nm}} \boldsymbol{p_{nm}}^H] \boldsymbol{w_{nm}}. \tag{3.44}$$

Beamforming fundamentals

The following beamforming fundamentals are first introduced in the space domain, and later extended to the spherical harmonics domain to account for spherical arrays. For the continuous description, a sound field on a sphere of radius r is considered, with sound pressure $p(k, r, \theta, \phi)$. The output of the beamformer y reads

$$y = \int_0^{2\pi}\int_0^{\pi} w^*(k, \theta, \phi) p(k, r, \theta, \phi) \sin\theta d\theta d\phi, \tag{3.45}$$

with $w^*(k, \theta, \phi)$ being the weighting function. For a microphone array, the discrete case is assumed, using Q microphones at position (r, θ_q, ϕ_q) and the respective sound pressures being described by $p_q(k) \equiv (k, r, \theta_q, \phi_q)$. They can be written in vector form, such as

$$\boldsymbol{p} = [p_1(k), p_2(k), \ldots, p_Q(k)]^T. \tag{3.46}$$

Accordingly, the spatial filter weights are also written in vector notation

$$\boldsymbol{w} = [w_1(k), w_2(k), \ldots, w_Q(k)]^T. \tag{3.47}$$

Weighting the pressure signals with the weighting function results in

$$y = \boldsymbol{w}^H \boldsymbol{p}. \tag{3.48}$$

In order to realize a beamformer with any desired property, for a given $\boldsymbol{p}$, the weights $\boldsymbol{w}$ have to be designed accordingly. This is commonly done assuming a single, unit-amplitude plane wave arriving from direction $\widetilde{\boldsymbol{k}} = (k, \theta_k \phi_k)$ which is sampled at microphone positions on a sphere $\boldsymbol{r} = (r, \theta_q, \phi_q)$. Here, the pressure vector $\boldsymbol{p}$ is replaced by the steering vector $\boldsymbol{v} = [v_1, v_2, \ldots, v_Q]^T$ representing the plane wave amplitude at microphone q, which is formulated by

$$v_q = e^{i\widetilde{\boldsymbol{k}} \cdot \boldsymbol{r}}, \quad 1 \leq q \leq Q. \tag{3.49}$$

This leads to the array output

$$y = \boldsymbol{w}^H \boldsymbol{v}, \tag{3.50}$$

defining the directional response of the array, with the steering vector being dependent on the used array configuration. For example, for rigid sphere arrays, the scattering effect is included in $\boldsymbol{v}$, as shown below in the general description of a spherical beamforming array by Eq. (3.54).

In the following, the pressure and weighting functions are represented in the spherical harmonics domain by $p_{nm}(k)$ and $w_{nm}(k)$ by which Eq. (3.45) can be extended, resulting in [218]

$$y = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} w_{nm}^*(k) p_{nm}(k, r). \tag{3.51}$$

Using matrix notation, this equation can be rewritten for an order-limited sound field as

$$y = \boldsymbol{w_{nm}}^H \boldsymbol{p_{nm}}, \tag{3.52}$$

or, according to Eq. (3.50), by [218]

$$y = \boldsymbol{w_{nm}}^H \boldsymbol{v_{nm}}. \tag{3.53}$$

Consequently, based on Eq. (3.8), a beamforming array in the spherical harmonics domain can be written as

$$p_{nm}(k, r) = v_{nm} = b_n(kr) \left[Y_n^m(\theta_k, \phi_k)\right]^*. \tag{3.54}$$

For beamforming, this equation can be applied to any array configuration through the mode strength $b_n(kr)$. The presented formulas also account for different sampling schemes which are represented by $\boldsymbol{S}$. In general form, this can be described by

$$y = \boldsymbol{w}^H \left[\boldsymbol{S}^H \boldsymbol{S}\right] \boldsymbol{p}. \tag{3.55}$$

In the following paragraphs, some beamforming approaches commonly employed in spherical microphone arrays are presented. The focus is on the PWD beamformer as it is the beamforming technique fundamental to the experiments conducted throughout this thesis.

Plane wave decomposition beamformer

A common, axis-symmetric beamformer is the so-called *plane wave decomposition* [220], or *regular* [154] beamformer. In [170], a beamformer was introduced providing axial symmetry around the look direction (θ_l, ϕ_l). It is described using the weighting function [218]

$$w_{nm}^*(k) = \frac{d_n(k)}{b_n(kr)} Y_n^m(\theta_l, \phi_l). \tag{3.56}$$

The beamforming weights are represented by d_n which are divided by b_n to eliminate the influence of the array such as sound scattering on a rigid surface. They control the beam pattern, i.e., the array response to a unit-amplitude plane wave. Substituting Eq. (3.56) in Eq. (3.52), leads to

$$y = \sum_{n=0}^{N} \frac{d_n(k)}{b_n(kr)} \sum_{m=-n}^{n} p_{nm}(k,r) Y_n^m(\theta_l, \phi_l), \tag{3.57}$$

which, for a beamformer after Eq. (3.54), leads to

$$y = \sum_{n=0}^{N} d_n(k) \frac{2n+1}{4\pi} P_n(\cos\Theta). \tag{3.58}$$

If the look direction of the beamformer is directly aiming at the arrival direction of the plane wave, (θ_k, ϕ_k), and the beamforming weights set to $d_n = 1$, then its output is written as

$$y = \sum_{n=0}^{N} \sum_{m=-n}^{n} w_{nm}^*(k) p_{nm}(k,r) = \sum_{n=0}^{N} \sum_{m=-n}^{n} a_{nm}(k) + Y_n^m(\theta_l, \phi_l). \tag{3.59}$$

Approximating the plane wave amplitude density function, the array output can be represented using plane wave components, as described by $y \approx a(k, \theta_l, \phi_l)$.

Although they are not specifically used in the experiments throughout this thesis, for reasons of completeness, the next section presents some optimal beam pattern designs, following the descriptions in [218].

Optimal beam pattern designs

The following beamformer designs are introduced to provide optimal beamforming performance for a specific property, like a maximum directivity or robustness against system uncertainty. Depending on the application, combined objective designs are also possible.

- **Delay-and-sum beamformer (DSB):** It is a common beamforming method in array systems of various geometries [114]. The beam steering is achieved by phase delays on the respective sensor signals. In [219], it was applied and investigated in spherical microphone arrays, and compared to phase-mode beamforming. The axis-symmetric beamforming weights for the DSB are given by $d_n(k) = |b_n(kr)|^2$. The beamformer output reads

$$y = \sum_{n=0}^{N} \sum_{m=-n}^{n} |b_n(kr)|^2 a_{nm}(k) Y_n^m(\theta_l, \phi_l). \tag{3.60}$$

  based on Eq. (3.51), and substituting the weights in Eq. (3.54) ad the measured sound pressure. DSB offers maximum WNG and enhances the SNR of the array, although the shortcoming of DSB are high side lobe levels. In order to compensate for this effect, the Dolph-Chebyshev beamformer was developed.

- **Maximum directivity beamformer:** This, also axis-symmetric, beamforming approach offers an improved array response into look direction based on an optimization approach for the DF. Note, that this beamforming property also holds for the PWD array. Therefore, a spherical harmonics domain formulation, with the coefficients being set constant, achieves the best directivity behavior, i.e., DI [218]. However, the maximum directivity is depending on the used order N. Consequently, to achieve a high directivity, a large number of microphones is needed.

- **Maximum WNG beamformer:** A maximum WNG beamformer achieves the highest WNG, providing the highest robustness of the array against sensor noise and other system uncertainties. The DSB beamformer, for example, offers maximum WNG properties.

- **Dolph-Chebyshev beamformer:** The Dolph-Chebyshev beam pattern achieves the lowest side-lobe level for a given array configuration [77]. It was applied for spherical arrays in [139].

- **Minimum variance distortionless response (MVDR) beamformer:** A more robust beamforming technique is the MDVR beamformer [67]. It is used for dereverberation and noise reduction in speech enhancement applications [110]. This design can specifically be tailored to the properties of the sound field, like noise due to a diffuse sound field. The adaptive version of the MDVR beamformer is the so-called *generalized side lobe canceler* [97].

- **Linearly constrained minimum variance (LCMV) beamformer:** The LCMV beamformer is an extension of the MVDR beamformer with additional constraints to the desired beam pattern. The so-called *null-constraint* accounts for disturbing sources in the noise field, allowing to control the width of the main lobe or the nulls in other directions.

For more detailed information on spherical array beamforming, refer to [279, 218].

3.1.5 Measurement Errors

This section addresses system errors in spherical microphone arrays which cannot be avoided in practical applications. Firstly, a theoretical description is presented based on an analysis framework developed in [215]. Secondly, the influence of such errors on the sound field representation is discussed based on simulation examples.

The subsequent error analysis comprises spatial aliasing[3], measurement noise, and microphone positioning offsets. In practice, two kinds of offset are possible: randomly distributed position errors and constant offsets. The latter are likely to arise in virtual scanning arrays due to inaccuracies in their mechanical construction, i.e., when the sensor is shifted to a certain direction. These constant offset errors, however, are not subject to the analyses described in Chapters 4 and 5 which only address the influence of randomly distributed offset errors. Furthermore, the directivity characteristic of the used microphones influences the array response. While in theory (which will be presented below) constant directivity patterns over the whole frequency bandwidth are assumed, in practice the directional response of microphones is frequency dependent. Consequently, the array response is distorted at respective frequencies. However, deviations in microphone directivity are not considered in the presented experiments and are therefore not addressed in the following descriptions. For more information on microphone directivity and constant offset errors affecting the array response, the reader is referred to [167].

[3] Note that spatial aliasing is more a system-inherent limitation than an actual measurement error.

Analysis framework

For analysis, a plane wave sound field coming from direction Ω_0 is sampled with a spherical array employing Q microphones. The plane wave is represented by $p_{nm} = b_n Y_n^{m^*}$ and the array by its weights $w_{nm}^* = d_n/b_n Y_n^m(\Omega_0)$ [215].

In order to analyze the effect of measurement errors, the array output is modelled as a signal contribution y_s and an error contribution y_{error}. Relating the power of y_s to the power of y_{error}, the effect of aliasing, transducer noise, and positioning errors can be analyzed for a given sampling scheme which is defined by the sets α_j and Ω_j for the respective microphone q_j. Each error contribution can be calculated as follows [215]:

$$E_a = \frac{|y_a|^2}{|y_s|^2} \qquad E_\Omega = \frac{|y_\Omega|^2}{|y_s|^2} \qquad E_e = \frac{|y_e|^2}{|y_s|^2}, \tag{3.61}$$

with E_a denoting the aliasing contribution, E_Ω the positioning error, and E_j the error due to uncorrelated noise. After [215], the signal power for a PWD beamforming array of order N is formulated by

$$\begin{aligned} |y_s|^2 &= \left| \sum_{n=0}^{N} d_n \sum_{m=-n}^{n} Y_n^m(\Omega_0) Y_n^{m^*}(\Omega_0) \right|^2 \\ &= \left| \sum_{n=0}^{N} d_n \frac{2n+1}{4\pi} \right|^2, \end{aligned} \tag{3.62}$$

which reduces to $(N+1)^4/(4\pi)^2$ with $d_n = 1$. Accordingly, the array output y can now be described as:

$$\begin{aligned} y = & \sum_{n=0}^{N} \sum_{m=-n}^{n} w_{nm}^* p_{nm} \\ & + \sum_{n=0}^{N} \sum_{m=-n}^{n} w_{nm}^* \underbrace{\left\{ \sum_{n'=N+1}^{\infty} \sum_{m'=-n'}^{n'} p_{n'm'} \epsilon_{a_{nmn'm'}} \right\}}_{y_a \text{ - aliasing error}} \\ & + \sum_{n=0}^{N} \sum_{m=-n}^{n} w_{nm}^* \underbrace{\left\{ \sum_{j=1}^{Q} \alpha_j e_j Y_n^{m^*}(\Omega_j) \right\}}_{y_e \text{ - noise error}} \\ & + \sum_{n=0}^{N} \sum_{m=-n}^{n} w_{nm}^* \underbrace{\left\{ \sum_{n'=0}^{\infty} \sum_{m'=-n'}^{n'} p_{n'm'} \epsilon_{\Omega_{nmn'm'}} \right\}}_{y_\Omega \text{ - positioning error}} . \end{aligned} \tag{3.63}$$

Spatial aliasing which is represented by the aliasing term $\epsilon_{a_{nmn'm'}}$ depends on the used sampling scheme. It is formulated by

$$\epsilon_{a_{nmn'm'}} = \sum_{j=1}^{M} \alpha_j Y_{n'}^{m'}(\Omega_j) Y_n^{m^*}(\Omega_j), \tag{3.64}$$

with $n \leq N$ and $n' > N$. The respective aliasing power can be computed with

$$|y_a|^2 = \left| \sum_{n=0}^{N} \sum_{m=-n}^{n} \sum_{n'=N+1}^{\infty} \sum_{m'=-n'}^{n'} d_n \frac{b_{n'}}{b_n} Y_{n'}^{m'^*}(\Omega_0) \times Y_n^m(\Omega_0) \epsilon_{a_{nmn'm'}} \right|^2, \tag{3.65}$$

which can be represented in a computationally more efficient form

$$|y_a|^2 = \left| \sum_{n=0}^{N} d_n \left\{ \sum_{n'=N+1}^{\infty} \frac{b_n'}{b_n} \frac{2n+1}{4\pi} \frac{2n'+1}{4\pi} \times \sum_{j=1}^{Q} \alpha_j P_n(\cos\Theta_j) P_{n'}(\cos\Theta_j) \right\} \right|^2. \tag{3.66}$$

The term in the curly brackets denotes the contribution of n to the overall aliasing error.

Under the assumption to be spatially uncorrelated, measurement noise is analyzed by adding the noise term e_j to the sound pressures $p(\Omega_j')$ which are measured at the respective position Ω_j'. With unit variance, the measurement noise at the array output is given by

$$\begin{aligned} E[|y_e|^2] &= \sum_{j=1}^{M} \alpha_j^2 |w(\Omega_j)| = \sum_{j=1}^{M} \alpha_j^2 \left| \sum_{n=0}^{N} \sum_{m=-n}^{n} \frac{d_n}{b_n} Y_n^m(\Omega_0) Y_n^{m^*}(\Omega_j) \right|^2 \\ &= \sum_{j=1}^{M} \alpha_j^2 \left| \sum_{n=0}^{N} \frac{d_n}{b_n} \frac{2n+1}{4\pi} P_n(\cos\Theta_j) \right|^2. \end{aligned} \tag{3.67}$$

Here, Θ_j is the angle between Ω_0 and Ω_j.

The positioning errors for a given sampling scheme are defined by

$$\epsilon_{\Omega_{nmn'm'}} = \sum_{j=1}^{M} \alpha_j [Y_{n'}^{m'}(\Omega_j') - Y_{n'}^{m'}(\Omega_j)] Y_n^{m^*}(\Omega_j), \tag{3.68}$$

with $n \leq N$ and $n' \geq N$. The formulation to calculate the power of the positioning error reads

$$|y_\Omega|^2 = \left| \sum_{n=0}^{N} \sum_{m=-n}^{n} \sum_{n'=N+1}^{\infty} \sum_{m'=-n'}^{n'} d_n \frac{b_{n'}}{b_n} Y_{n'}^{m'^*}(\Omega_0) \times Y_n^m(\Omega_0) \epsilon_{\Omega_{nmn'm'}} \right|^2, \tag{3.69}$$

which can be represented in a computationally more efficient form:

$$|y_\Omega|^2 = \left| \sum_{n=0}^{N} d_n \Big\{ \sum_{n'=N+1}^{\infty} \frac{b'_n}{b_n} \frac{2n+1}{4\pi} \frac{2n'+1}{4\pi} \times \sum_{j=1}^{Q} \alpha_j P_n(\cos\Theta_j)[P_{n'}(\cos\Theta'_j - P_{n'}(\cos\Theta_j)] \Big\} \right|^2 . \tag{3.70}$$

In this equation, Θ'_j represents the angle between Ω_0 and the actual, erroneous sensor position. This deviation is denoted by Δ which is added to the respective angles, resulting in actual positions for each microphone $\theta'_j = \theta_j \pm \Delta$ and $\phi'_j = \phi_j \pm \Delta$. Based on this framework, some simulation examples are given in the following.

Simulation examples

In order to illustrate the influence of measurement errors on the array performance, some simulation examples are presented next, based on the formula provided above. Following [215], a plane wave sound field coming from direction $\Omega(\theta, \phi) = (25.7°, 60°)$ is considered with the array looking in the same direction. Exemplarily, Lebedev and Gaussian sampling schemes are employed and compared regarding their error robustness using sensor positions and array weights from [29]. Table 3.1 lists the array configurations used in the simulations. Arrays 1 and 2 are chosen to show the error influence on different sampling schemes using arrays with similar N_{max} while arrays 3 and 4 provide nearly the same number of microphones Q for different N_{max}. For each array, an effective order of $N = 2$ is calculated. The variance of the input measurement noise is denoted by σ^2. Random sensor positioning errors are simulated with a uniformly distributed offset Δ which is added to all samples. The maximum deviation is set to 0.3° which approximately equals to a maximum offset of 1 mm for an array with a radius of 20 cm. Also note that random positioning errors are calculated and averaged over 50 iterations.

Table 3.1: Array configurations used to illustrate error contributions. Q is the number of samples, with N_{max} being the respective maximum order, and N the effective order used in the calculations.

Array #	Sampling Quadrature	Q	N	N_{max}	σ^2	$\|\Delta_{max}\|$
1	Lebedev grid	14	2	2	1	$\approx 0.3°$
2	Gauss grid	18	2	2	1	$\approx 0.3°$
3	Lebedev grid	38	2	4	1	$\approx 0.3°$
4	Gauss grid	36	2	3	1	$\approx 0.3°$

For all array configurations listed in Table 3.1, the respective error contributions are shown in Figure 3.5 as a function of kr (in dB).

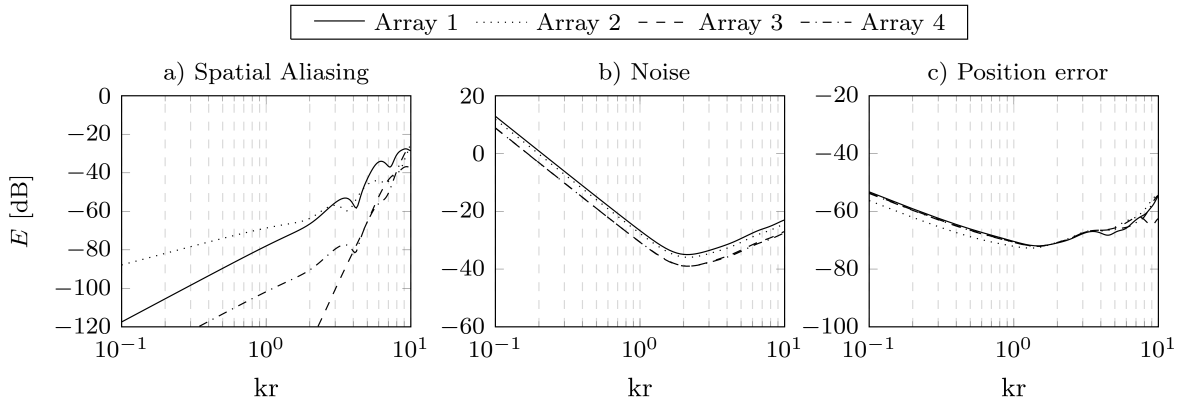


Figure 3.5: Error contributions E as a function of kr (in [dB]) for spatial aliasing, measurement noise, and microphone positioning errors, according to the array designs from Table 3.1.

Spatial aliasing: Due to the discrete nature of sampling, spatial aliasing cannot be avoided in practice. It depends on the number of samples used, their geometrical distribution on the surface of a sphere and the transform order N for sound field decomposition. The sphere's radius r defines the respective frequency limit. Spatial aliasing corrupts the directional response of the array at high kr, i.e., in the high frequency range, as shown in Plot a) in Figure 3.5. Here, the Lebedev quadrature provides the highest robustness against spatial aliasing for a given order N_{max} which can be seen in the plot. If the sound field satisfies $kr < N_{max}$, then spatial aliasing should be negligible. Arrays 1 and 3, represented by the solid and dashed line, show contributions below -60 dB and -80 dB for $N = kr$, respectively. Gaussian sampling on the other hand seems more prone to spatial aliasing. This behavior can clearly be seen comparing arrays 3 and 4. Also, using the Lebedev grid for sampling requires the lowest amount of samples for a given sound field order N compared to all other quadratures. It is therefore the sampling scheme of choice in practical applications. In addition, spatial aliasing is not depending on the sound arrival direction. Perceptual investigations in [8] showed that spatial aliasing leads to high-frequency *ringing* sounds corrupting the auralization quality. Some methods for aliasing reduction, like applying spatial low pass filters before sampling when operating the array at high kr, are proposed in [222]. Recently, an approach for optimal aliasing cancellation in spherical array beamforming was proposed in [6].

Measurement noise: Although todays microphones provide high SNRs of over 90 dB, microphone noise cannot be disregarded when applying a PWD on the sound field. As depicted in plot b) in Figure 3.5, transducer noise can lead to extreme low-frequency amplifications when calculating higher orders at low kr. This is due to numerical instabilities of the Bessel functions j_n becoming (nearly) zero and leading to a division by a very small number in the calculation of the array weights. The highest robustness against noise can be found at $N \approx \text{kr} \approx 2$, which rapidly decays for higher and lower kr. However, the robustness can be increased using more microphones for sampling, as can be seen comparing arrays 1 and 2. Also, it becomes clear that the influence of measurement noise is marginally depending on the used sampling scheme. In addition, as shown in [167], the sound direction of arrival has no influence. Recall, that this error is the reciprocal of the WNG. In [216], an approach to recover the nulls in the Bessel-functions was proposed for optimization. Another approach to reduce the influence of noise is to apply radial filters for smoothed amplification limitation as described in [30].

Microphone positioning errors: The error performance of microphone positioning offsets is similar for all quadratures, as can be seen in plot c) in Figure 3.5. For all arrays, the error is below -50 dB, whereas an error minimum can be found at around $N \approx \text{kr} \approx 2$. Also note that the behavior of positioning errors is quite similar to measurement noise or its inverse, the WNG. The error contribution for positioning errors rises for lower and higher kr, although noise has a stronger impact on low kr. In addition, the impact of positioning errors is independent of the DOA.

3.1.6 Head-Related Transfer Functions

In the scope of this thesis, the auralization is done using non-individual binaural synthesis, based on a spherical set of HRTFs measured with a dummy head. In the following, the HRTFs are also described in the spherical harmonics domain, following [81]. The sound pressure measured at the left ear is represented in spherical coordinates [221]

$$p^l(k) = \int_0^{2\pi} \int_0^{\pi} a(k,\theta,\phi) H^l(k,\theta,\phi) \sin\theta d\theta d\phi, \tag{3.71}$$

where $a(k,\theta,\phi)$ is the complex amplitude of a plane wave coming from direction (θ,ϕ) and $H^l(k,\theta,\phi)$ being the respective HRTF of the left ear. The definition for the right ear is similar using the right ear HRTF $H^r(k,\theta,\phi)$. The pressure functions can be represented for the left ear [221]

$$p^l(k) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \bar{a}_{nm}^*(k) H_{nm}^l(k). \tag{3.72}$$

The right ear representation is similar using the right ear HRTF $H_{nm}^r(k)$, where $\bar{a}_{nm}^*(k)$ is the spherical harmonics representation of $a(k, \theta, \phi)^*$. Following (3.72), the spherical harmonics representations of the right and left HRTF are given by $H_{nm}^r(k)$ and $H_{nm}^l(k)$, respectively. Using a spherical microphone array for sound capturing and applying a PWD in the spherical harmonics domain yields the spherical Fourier coefficients $a_{nm}(k, r, \theta, \phi)$. Since in real measurement situations the number of microphones employed in the array is limited, only a limited order of spherical harmonics can be computed, leading to the following pressure representations at both ears:

$$p^{l/r}(k) = \sum_{n=0}^{N} \sum_{m=-n}^{n} \bar{a}_{nm}^*(k) H_{nm}^{l/r}(k) \tag{3.73}$$

In the presented experiments, the array data and the HRTFs are combined in the spherical harmonics domain for binaural synthesis [73].

3.2 Perceptual Fundamentals

This section addresses the perceptual fundamentals related to this thesis. The first section is about the human auditory system and its signal processing stages. In this context, three specific concepts are presented: the threshold of hearing, masking, and critical bands. The second section outlines the concept of binaural hearing and the third describes the perception of spaciousness in terms of ASW and LEV. The final section deals with auditory models, with a special focus on the PEAQ model as it is the auditory model used in the experiments documented later in this thesis.

3.2.1 Human Auditory System

The human auditory system comprises multiple sound processing stages transforming incoming sound waves into a basilar membrane representation. Figure 3.6 shows the processing stages for an incoming sound wave, following [304] and [98]. Basically, the auditory system is divided into three parts: The outer, middle, and the inner ear. The first, comprising pinna and ear canal, is used for protection, amplification, and the localization of sound. An incoming sound wave is collected by the pinna and sent

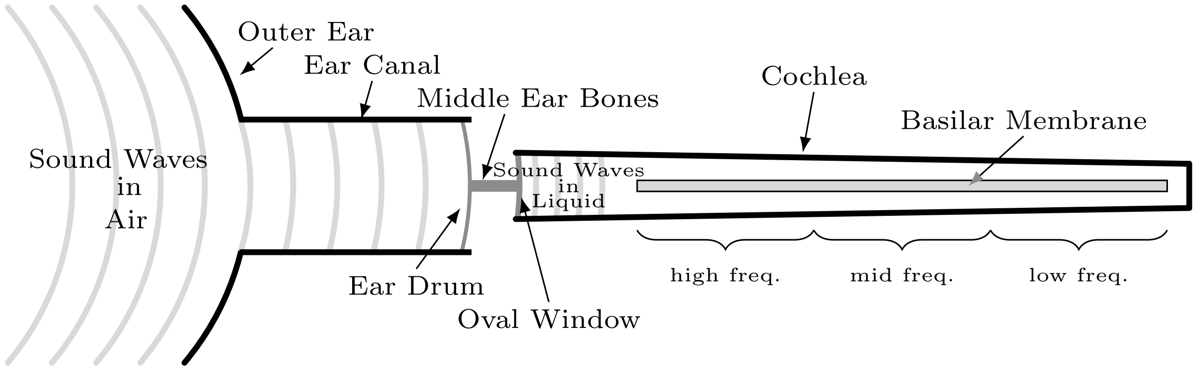


Figure 3.6: Processing stages of the human auditory system (after [258]).

through the ear canal to the ear drum (or tympanic membrane) which separates the outer and middle ear sections. The second part, the middle ear, transforms the induced air vibrations into mechanical ones by passing the vibratory pattern from the ear drum to the ossicular chain which is represented by the middle ear bones malleus, incus, and stapes[4]. Besides protection, the main function of the middle ear is impedance matching (from mechanical to fluid transmission in the inner ear) as well as selective oval window stimulation and pressure equalization using the Eustachian tube. The third section, the inner ear, is a fluid-filled system called cochlea. The cochlea is spiral shaped and is excited via the oval window membrane by vibrations from the stapes of the ossicular chain. A main structural element is the basilar membrane, carrying hair cells which interact with the auditory nerve by means of synaptic junctions. Here, the fluid movements which cause the basilar membrane to move, and consequently the hair cells, are translated into neural responses. They pass the information to the central auditory nervous system as electrochemical (neural) signals for information processing in the brain.

While the described signal processing stages presented the physiology of the human auditory system, psychoacoustics deals with the perception of sound, a *blackbox* described by curves derived experimentally. Respectively, the basic psychoacoustic principles considered important in the context of this thesis are reviewed in the following sections, specifically addressing the threshold in quiet, masking in the time and frequency domain, perceptual bands as well as the fundamentals of spatial hearing—and the perception of spaciousness (in terms of concert hall acoustics). For further information on human auditory perception and the corresponding psychoacoustic fundamentals, the reader is referred to the respective literature, e.g., [310, 311, 35].

[4] Note that sound can also be transmitted directly to the inner ear based on bone transmission via the skull, hence, bypassing the middle ear.

Absolute threshold of hearing

The absolute threshold of hearing, or threshold in quiet, describes the sound pressure level required for a specific frequency to be perceived by a specific listener. The threshold was determined experimentally, providing high reproducibility of ±3 dB per subject [88, 310]. A rough approximation of this threshold is shown in Figure 3.7 (after [274]). However, the curve would stronly vary between subjects.

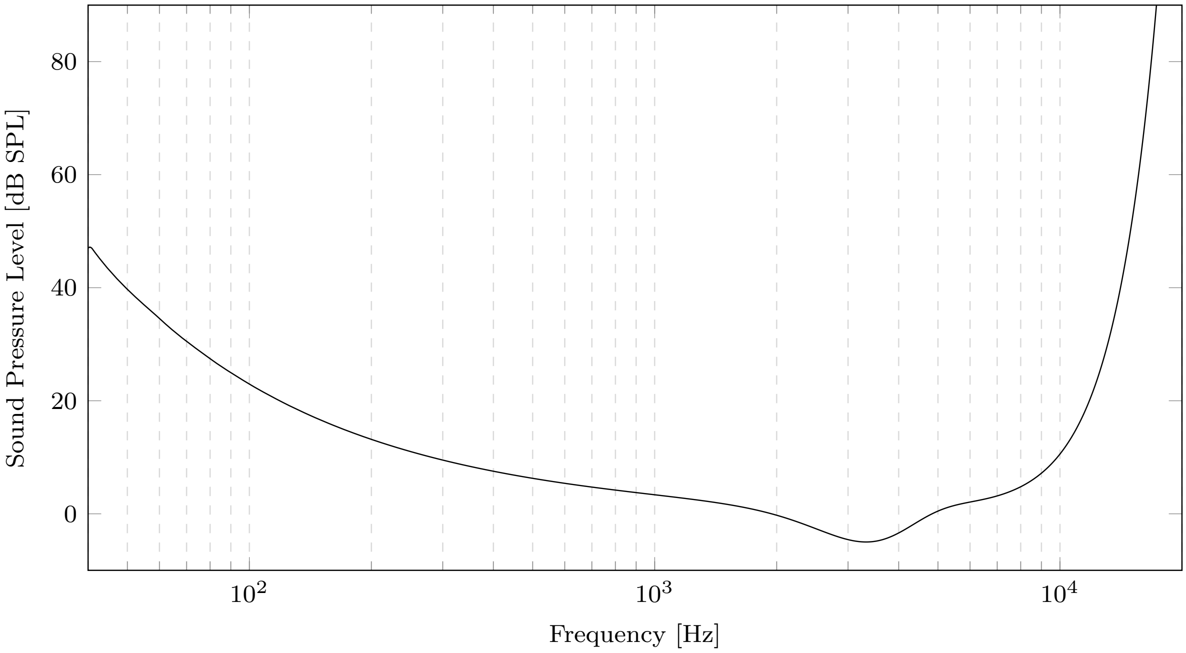


Figure 3.7: Threshold in quiet (after [274]).

The threshold in quiet is affected by the superposition of a stimulus with internal noise (induced by blood flow) in the low frequency range as well as the outer and middle ear transfer function at high frequencies [274]. The resonance of the ear canal around 3 kHz indicates an increased sensitivity of the auditory system. The threshold in quiet is representative for a young listener with accurate hearing and most likely differs for each individual due to, among other reasons, age or potential damage.

Masking

Masking describes the effect when a sound becomes inaudible due to the presence of another sound. This is the case when a masker excites the basilar membrane at a specific location in a way that a weaker signal in the vicinity of that location can not be detected. The masking threshold depends on the temporal and spectral structure of masker and maskee. The descriptions in this section follow [310, 264, 111].

Masking can basically be divided into simultaneous and non-simultaneous masking. Simultaneous masking is determined in the frequency domain based on the shapes of the magnitude spectra of masker and maskee while non-simultaneous masking is separated into backward and forward masking. Four types of frequency masking can be modelled, namely *Noise-Masking Noise* (NMN) [111], *Tone-Masking Noise* (TMN) [118], *Noise-Masking Tone* (NMT) [244], and *Tone-Masking Tone*.

Note that all curves presented in this section were derived experimentally based on human listeners. However, some formula exist in the literature to roughly approximate these curves [310].

Pure tones masked by broadband noise: A white noise signal, i.e. with a constant spectral density over all frequencies, is used to mask a pure tone. Figure 3.8 shows the level of the test tone as a function of frequency. The solid lines indicate the level of the pure tone and the dotted line the threshold in quiet. Note that the line is horizontal at low frequencies and ascending for frequencies above 500 Hz. The masking thresholds rise approximately with a 10 dB slope per decade. When the density level is increased by 10 dB then the masking thresholds also rise the same 10 dB, indicating the linear masking behavior of broadband noise.

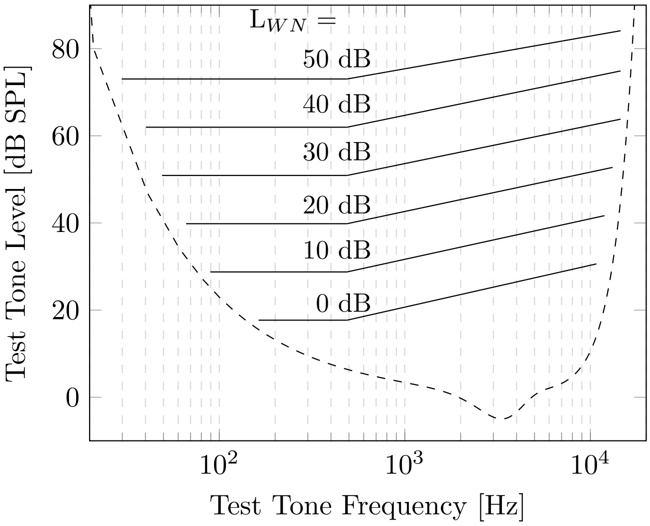


Figure 3.8: Level of the test tone masked by wide band white noise (after [310]).

Pure tones masked by narrowband noise: While the previous paragraph addressed masking by broadband noise, this paragraph describes masking when a pure tone is masked by narrowband noise.

The plots in Figure 3.9 show the masking thresholds of a pure tone masked by white noise with critical bandwidth.

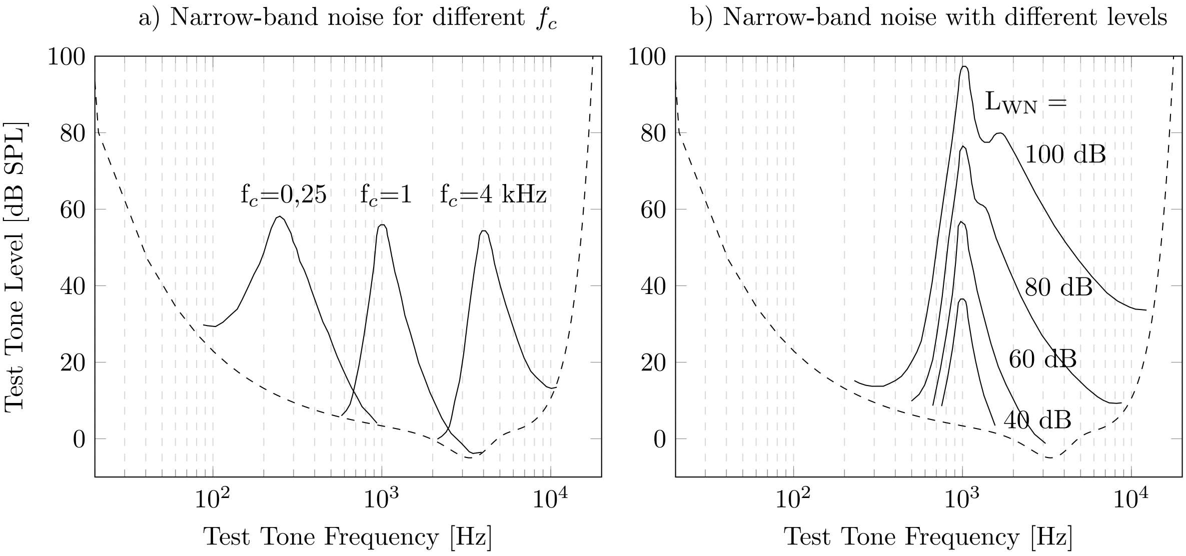


Figure 3.9: Masking thresholds (solid lines) for pure tones masked by white noise with critical bandwidths for different center frequencies in plot a) and different levels in plot b) (after [310]).

The dashed line indicates the threshold in quiet as defined by Eq. (3.2.1). Plot a) shows the masking threshold for pure tones masked by band limited white noise at center frequencies $f_c = 250\,Hz$, $f_c = 1000\,Hz$, and $f_c = 4000\,Hz$ at a constant level of $L_{WN} = 60\,dB$. The respective bandwidths are 100 Hz, 160 Hz, and 700 Hz. Plot b) illustrates the dependence of the masking threshold for different noise levels, i.e., $L_{WN} = 40\,dB$, 60 dB, 80 dB, and 100 dB for a center frequency of $f_c = 1000\,Hz$. Note the level dependence of the masking threshold in plot b) in Figure 3.9 and the dips for levels $L_{WN} = 80\,dB$ and $L_{WN} = 100\,dB$. They can be explained by the nonlinear behavior of the auditory system. The rising slope below the f_c of the masker is steep and independent of the noise level while the falling slope at high frequencies shows a more smoothing decay with increasing level.

Next, the influence of high- or low-pass filtered noise is addressed. Figure 3.10 presents the masking threshold of a low pass and high pass noise with levels of $L_{WN} = 0$, 20, and 40 dB. The first is shown in plot a) with a corresponding cut-off frequency of $f = 900\,Hz$, and the latter in plot b) with $f = 1100\,Hz$, respectively. Note that the slopes show a similar behavior for low and high frequencies as the full band white noise case in Figure 3.8.

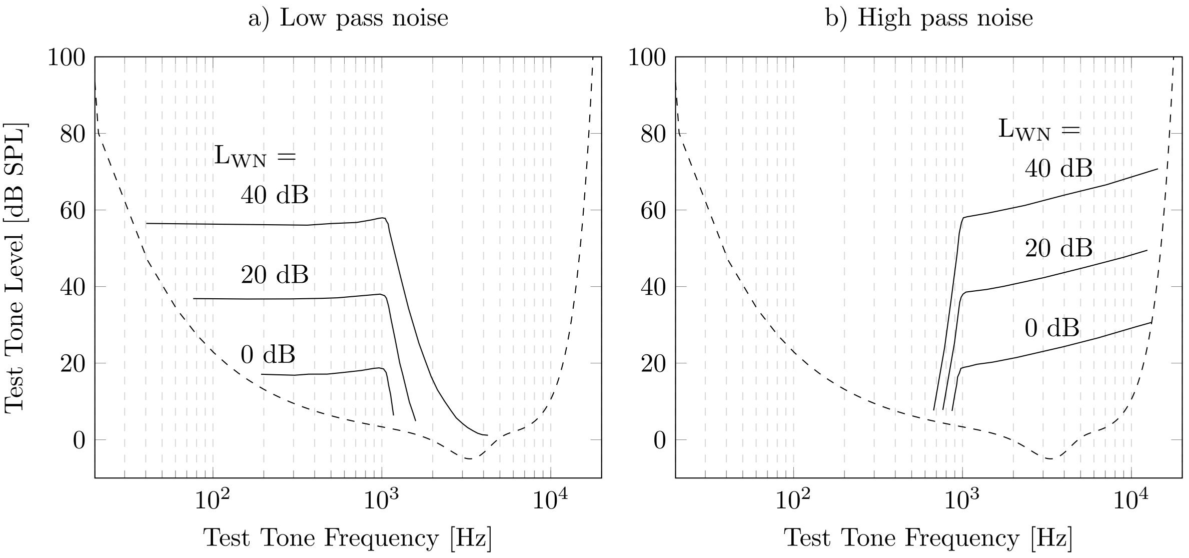


Figure 3.10: Masking thresholds (solid lines) for pure tones masked by band-limited noise of different levels with a low pass (plot a)), and a high pass characteristic in plot b) (after [310]).

Pure tones masked by pure and complex tones: Masking also occurs when pure tones, i.e., sinusoidal tones, are masked by other pure or complex tones. The latter represent a sound composed of a fundamental frequency and several harmonics, i.e, integer multiple of the fundamental frequency. Figure 3.11 shows both phenomena.

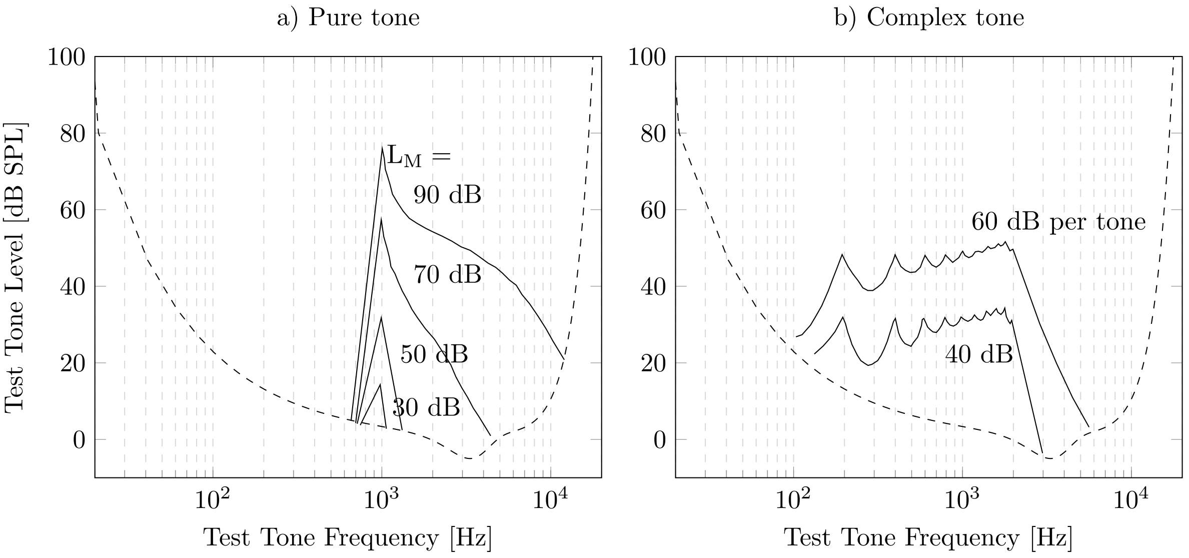


Figure 3.11: Principle of pure tones masked by pure tones in plot a) and pure tones masked by complex tones in plot b) (after [310]).

In [310], a 1 kHz tone at a level of 80 dB was used as masker, measuring Two effects: first, the dominant perception of beating which occurs when the test frequency is in the vicinity of the test tone. Second, the perception of an additional tone, the so-called difference tone, at test tone frequencies near 1.4 kHz. This is due to nonlinear distortions in the hearing system, and only experienced assessors were able to distinguish between the threshold of the difference tone and the maskee. Similar effects were observed when pure or complex tones tones are masking the noise signal.

Time domain masking: In the previously described masking effects, masker, and maskee were presented simultaneously for a long time, whereas time domain masking (or temporal masking) occurs before and after a signal when a masker is present. The basic principle of temporal masking is shown in Figure 3.12.

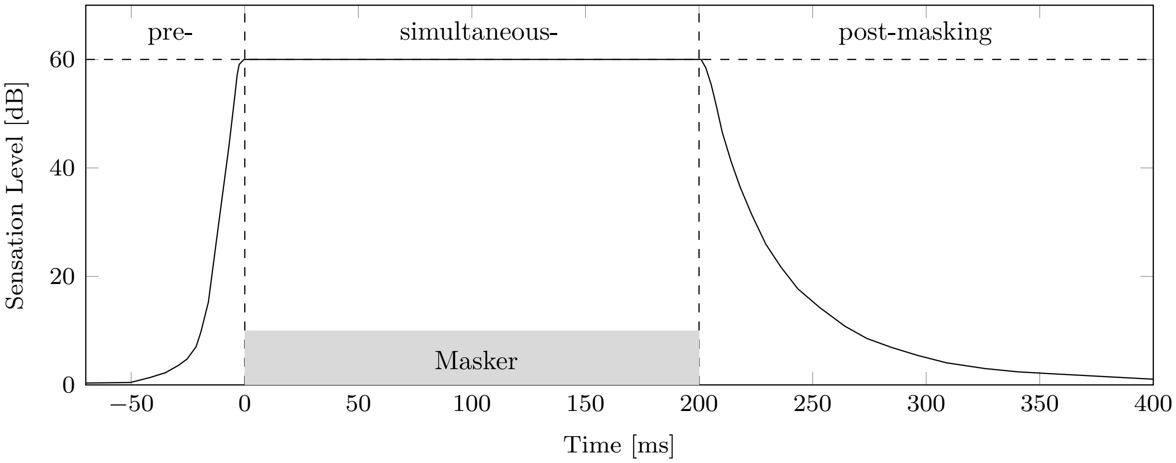


Figure 3.12: Principle of temporal masking (after [310]).

Temporal masking can be separated into three regions: premasking, simultaneous masking, and postmasking. Premasking represents the build-up time for a masking signal. While early research indicated premasking lengths of 5–20 ms [310], later experiments found that premasking is as short as about 1.5 ms [262]. Then, for a certain time period, masker and maskee are presented simultaneously. When the masker is turned off, postmasking occurs which depends on the masker duration. After a short delay of about 5 ms, it takes approximately 200 ms until the sensation level reaches the threshold in quiet. But also these values strongly vary for different signal types, as postmasking of impulses can be as short as about 1.5 ms [262].

The masking phenomena described in this section determine how frequencies are grouped by the human auditory system. Such frequency groups are called critical bands, an important psychoacoustic concept which will be described next.

Critical bands

Critical bands, as used here, are perceptually-motivated frequency groups representing the frequency-to-place transformation in the cochlea. Specifically, a critical band indicates the *effective* bandwidth of a white noise signal that masks a tone, assuming that the tone is masked by that specific portion of the noise being in the vicinity of the tone's frequency [88]. Accordingly, the tone is masked, if its intensity is smaller as the noise portion's intensity within this frequency group. Based on the width of these critical bands, which are determined experimentally, different band-representations exist. Common scales, representing this frequency-pitch relationship, are the Bark, Mel, and Equivalent Rectangular Bandwidth (ERB) scale. Figure 3.13 shows the relation between these scales and their position on the basilar membrane (after [311]).

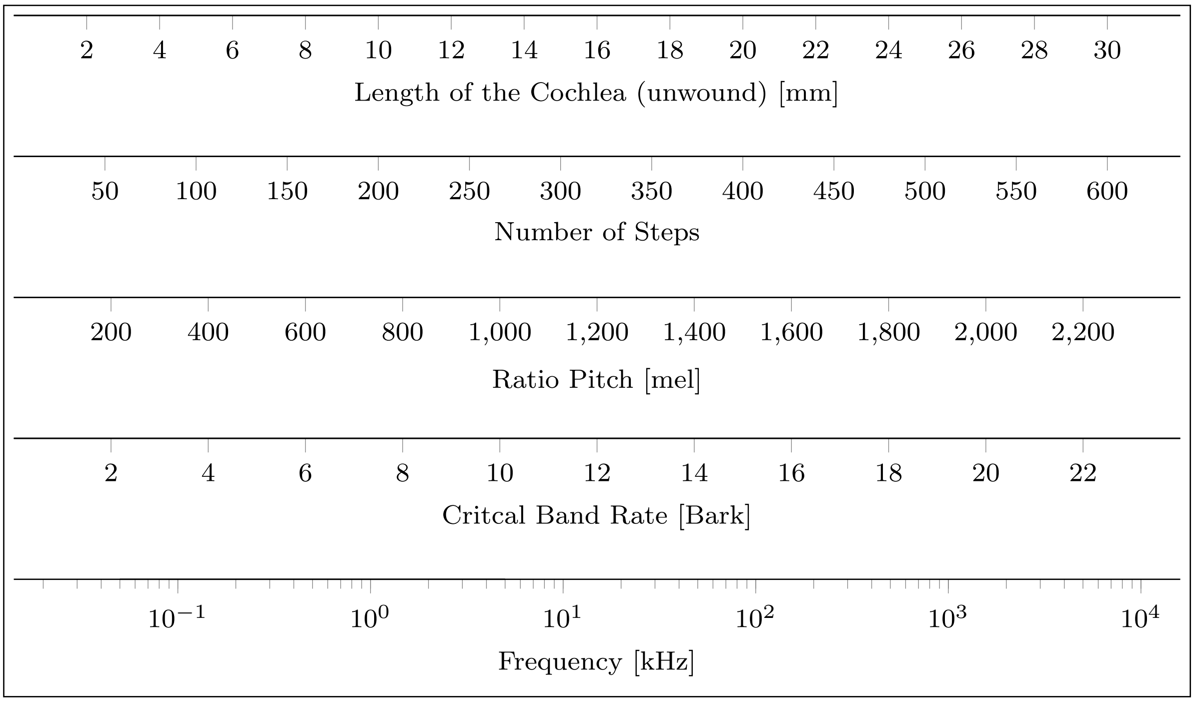


Figure 3.13: Relation of critical bands to cochlea and frequency (after [310]).

If sound reaches the listener from a certain direction, then the sound waves are preprocessed by the peripheral hearing system (body, head, pinna, and ear canal) before reaching the ear drum. As described in the introduction of Section 3.2.1, these waves are further transferred to the cochlea. The induced waves are traveling along the length of the basilar membrane which is connected to neural receptors. Depending on their positions on the cochlea, a frequency-place transform takes place leading to peak responses at frequency-specific membrane positions.

In practical applications, like in audio coding, the frequency grouping of the auditory system can be modeled using a filter bank with overlapping bandpass filters with different frequency-depending bandwidths. An alternative representation is the Bark scale. For a Bark representation, the frequency axis from 0 to 16 kHz is separated into 24 bands with adjacent slopes. The Bark scale is a linear scale for tonality, so a doubling of Bark yields a perceived doubling of tonality. The relation between Bark and frequency follows a linear behavior below 500 Hz, and is logarithmic for frequencies above [311]. A related representation is the Mel scale which connects to Bark in a way that 1 Bark = 100 mel [311], also providing a linear pitch scale along the basilar membrane. A third frequency representation is the ERB scale which is similar to the Bark scale [182]. Note that all these scales were determined experimentally using human listeners. However, some formula exist in the literature to roughly approximate these scales [310].

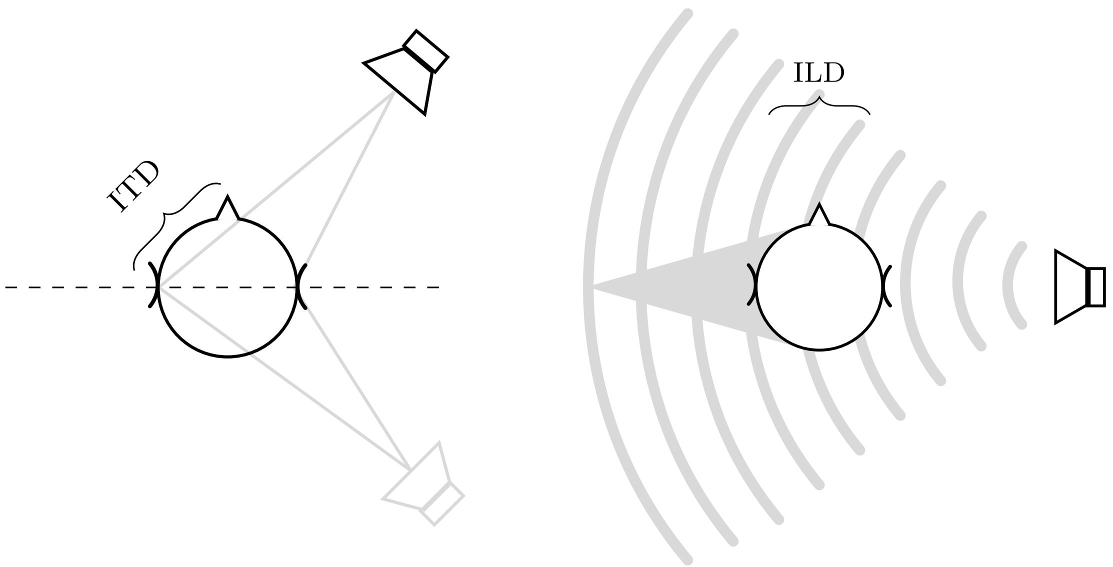
3.2.2 Binaural Hearing

Binaural hearing mainly involves the principles for sound source localization but it is also related to the perception of spaciousness, as described in Section 3.2.3. Although the brain can determine the location of a source based on monaural cues only [207], in binaural hearing, mainly the differences in amplitude, delay or phase, and spectral characteristics between both ear signals are evaluated. The respective cues are the Interaural Time and Level Differences, ITD and ILD, as well as the correlations between both ear signals which are expressed by the Interaural Correlation (IC). Based on the theoretical fundamentals provided by Lord Rayleigh, the principles for binaural hearing were summed up as the so-called *Duplex Theory* in [271]. These cues play an important role in sound source localization and the perception of spaciousness.

Source localization is frequency dependent with the transition frequency at approximately 1.5 kHz, which corresponds to wavelengths in the approximate size of a human head. In the frequency range below 1.5 kHz, phase differences are evaluated, whereas amplitude differences account for localization at frequencies above 1.5 kHz [35, 234]. So higher frequencies with shorter wavelengths are reflected at the boundary, hence, leading to interaural amplitude differences due to the head acting as sound barrier, whereas for lower frequencies, i.e., longer wavelengths, the sound wave is bended around the head, resulting in interaural time or phase differences. In addition, reverberation as well as the temporal and spectral structure of the signal also influence spatial perception and source localization. For more details on binaural hearing and psychoacoustics, see [35, 310, 234, 251, 181].

Interaural time and level differences

When the wave front of a sound source from a certain direction arrives at the head of a listener, then it reaches the source-facing ear first and, depending on head's size and shape, the averted ear a few milliseconds later[5]. This difference is the ITD, which depends on the direction of the sound source. It has a maximum for a source in the lateral direction, and a minimum, i.e., no delay at all, for a source directly in front of the listener. Note that for sound sources positioned in the sagittal plane, the same ITDs are provoked. This also holds for source positions on conical surfaces in the frontal plane, as illustrated by the grayed loudspeaker symbol in Figure 3.14 a). These regions, producing similar ITDs and ILDs, are the so-called *cones of confusion*, leading to localization ambiguities in binaural synthesis. In real life, these ambiguities are compensated for by small head movements. Figure 3.14 b) shows the principle of source localization based on ITDs.



(a) ITD for a sound source in the frontal right. The grayed source shows location ambiguity.

(b) ILD for a sound source on the right side of a listener. Gray area indicates the head shadow.

Figure 3.14: Principles of ITD in plot a) and ILD in plot b) (after [157]).

The second cue ILD describes the amplitude differences in the spectrum between both ear signals due to the head acting as a sound barrier. It is indicated by the gray area, the head-shadow, in plot b) in Figure 3.14 for a sound source on the right hand side of a listener. For all source positions on the sagittal plane, the ILDs become zero.

[5] Note that the term Interaural Phase Differences (IPD) is also often used. The IPDs are equivalent to the ITDs.

Interaural correlation

The third cue in binaural hearing is the Interaural Correlation (IC) which is represented by the cross correlation of the pressure signals at both ears, resulting in the Interaural Cross Correlation Coefficient (IACC). The IACC is the maximum of the Interaural Cross Correlation Function (IACF) which measures the similarity between both ear signals in a sound field. The IACF can be calculated (after [1])

$$\mathrm{IACF}(\tau) = \frac{\int_{t1}^{t2} x(t)y(t+\tau)dt}{\sqrt{\int_{t1}^{t2} x^2(t)dt \int_{t1}^{t2} y^2(t)dt}}, \tag{3.74}$$

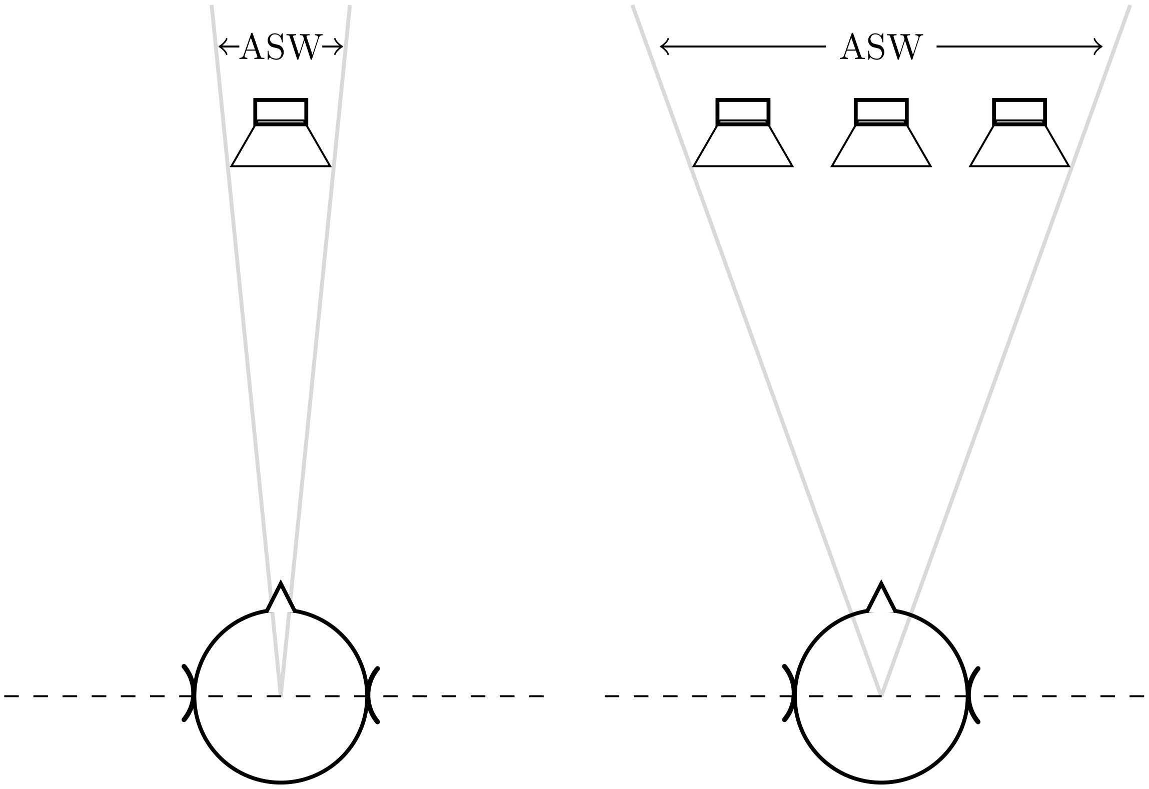
with $x(t)$ and $y(t)$ being the left and right ear signal, respectively, and τ the temporal offset between both signals. The IACF has been identified as a measure for the perception of spaciousness or spatial impression [120]. In particular, a low correlation corresponds to a greater spatial impression. Early research used the IACF to model spatial perception evaluating the ITDs [128], or to describe the degree of binaural fusion of multiple sound sources which was found at the maximum of the IACF [242]. In [268], an overview of Just-Noticeable Differences (JND) in ITD, ILD, and IC is given.

3.2.3 Perception of Spaciousness

The following descriptions address the perception, or impression of space, or spaciousness as it is also called. It occurs when sound is coming from multiple (or from all) directions. The impression of spaciousness strongly depends on the reflection properties of the environment but also other effects contribute such as, for example, the precedence effect [68, 183].

As described earlier in the state of research in Section 2.1, the perception of spaciousness refers to an acoustic room impression which, in reality, is a multi-dimensional percept. Over decades, a number of attributes and predictors for spaciousness were introduced and are, until today, still used in quality evaluations of spatial room acoustics, like in concert halls. Although recent research proposed further quality features [292], specifically two metrics have been agreed upon which are strongly related to the perception of spaciousness in rooms and therefore spatial quality, namely ASW and LEV. Although these features are connected, both can be perceived as two distinct senses [184]. The properties of these two attributes are described next in more detail, since both are used as quality attributes in the evaluations in Chapter 4. For more information, a comprehensive overview on spatial room impression can be found in [54], specifically addressing ASW and LEV.

Apparent source width: ASW (also auditory scene width) is a measure for spaciousness which describes the perceived broadening effect of a sound source appearing wider than its visual or physical size [133]. A schematic of ASW perception is shown in Figure 3.15 for a narrow sound source in plot a) and a source with wider extent in plot b), both positioned directly in front of a listener.



(a) ASW for a narrow sound source

(b) ASW for a wide sound source

Figure 3.15: ASW for a narrow sound source in front of a listener in plot a) and a wider source in plot b) (after [234]).

The perception of ASW is influenced by a number of factors. The following list presents some important properties, addressing predictors for ASW and the influence of specific sound field characteristics on its perception.

- ASW depends on lateral and rear reflections in the early sound field leading to a decorrelation at both ears and the perception of spaciousness [78].
- ASW is related to loudness. For increasing sound levels, also the perceived broadening effect of the source, i.e., ASW, increases [12].

- Distance perception is connected to ASW as well. It can be predicted assessing the early sound strength G$_e$ [151].
- Mainly low frequencies (< 500 Hz) contribute to ASW perception, which can be quantified by evaluating the sound strength at these frequencies, i.e., G_{low} [197].
- ASW strongly correlates with localization accuracy [283]. By definition, an increased ASW coincides with increased localization uncertainty.
- Also, the presented signal type has a major influence on ASW perception [283].
- The lateral fraction (LF), in particular the early lateral fraction LF$_e$, is a predictor for ASW [38] (LF$_l$, the late lateral fraction, is used for LEV prediction).
- ASW can be predicted by evaluating the IACCs of the early sound field in every time-frequency bin [137].

In addition, also further measures for ASW prediction have been introduced over the years, like DSB. For more information, please refer to the state of research in Section 3.2, or the respective literature [166] or [54].

Listener envelopment: LEV strongly relates to the perception of spaciousness and describes the feeling of being *inside* or *surrounded* by sound. Figure 3.16 shows a schematic of LEV for a listener *being in* the sound field, i.e., the acoustical scene.

Some results from the literature addressing important aspects of LEV perception are listed in the following:

- LEV is mainly related to the late reverberating part of the sound field [39, 95].
- However, also early reflections have been found to contribute to LEV [38, 195].
- LEV can be predicted by evaluating the late part of the sound field in every time-frequency bin [137].
- LEV is influenced by reflections from different directions, like from above, behind, and also frontal reflections can increase the feeling of LEV [184, 95, 82].
- The late lateral sound field has high correlation to LEV [95, 82]
- Distance perception is related to LEV [151], which could be predicted using the late G$_l$ and BF-ratio in the late sound field.

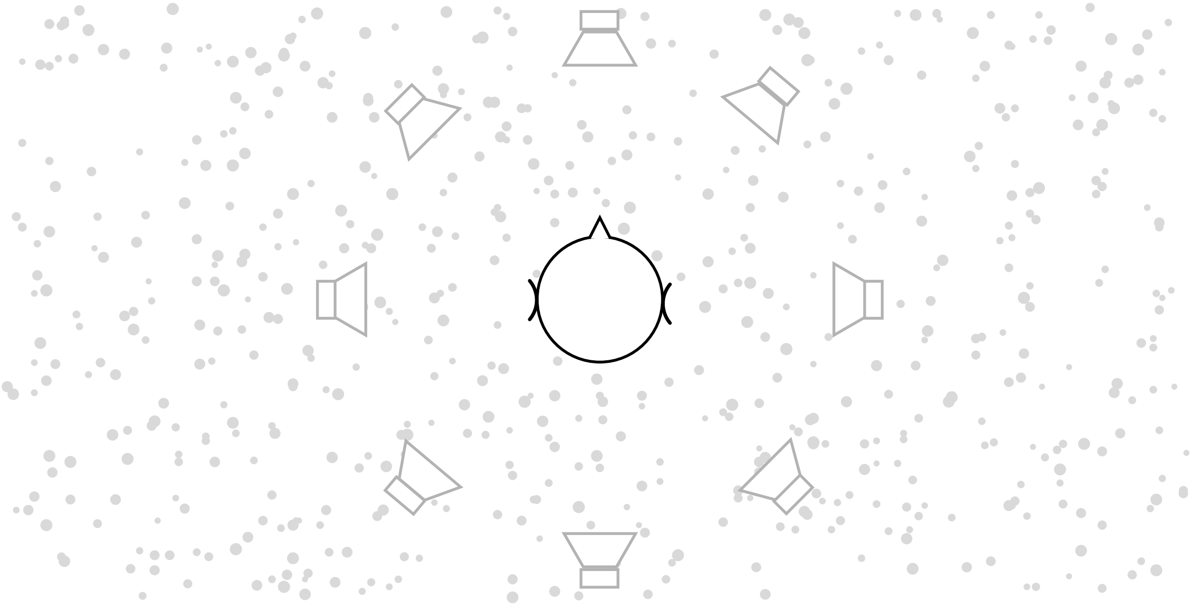


Figure 3.16: Schematic of LEV, i.e., sound surrounding a listener (after [184]).

Although ASW and LEV can be perceived as two distinct senses, it has been shown that they influence each other in a way that an increased ASW also leads to a slightly increased LEV perception and vice versa [38, 195].

3.2.4 Auditory Models

Early auditory models have been introduced in speech coding research in the late 1970s to assess the coding quality, like [246]. In the following years, research on auditory modeling was intensified, resulting in a variety of improved models of auditory perception [132, 41, 16, 275]. Today, such models are employed in a wide range of applications, like in audiology for assessing disorders in binaural listening, hearing aids, room acoustics as an echo detector, or in robot audition [34]. Recently, the auditory modeling toolbox (AMT), an open source toolbox for auditory modeling in MATLAB or Octave was presented in [259, 164], provided by the Aural Assessment By means of Binaural Algorithms (AABBA) initiative [36]. AMT comprises multiple model types such as peripheral models estimating the direction-continuous time of arrival [307], spatial models for source localization [13] as well as models for speech intelligibility prediction in reverberant environments [145]. A model for ASW and LEV prediction, based on the evaluation of IACCs in the early and late sound field, respectively, was developed and evaluated in [137]. Recent research also takes higher cognitive processes into account, like the model proposed in [212, 214]. In the following, PEAQ is described in more detail because it is used in Chapter 5.

PEAQ model

PEAQ was originally designed to predict the quality of an audio coder in terms of the induced coding artifacts [276], and was standardized as ITU-R.BS.1387 [226]. Modeling the sensory processing stages of human auditory perception, it allowed to measure these artifacts by evaluating the spectro-temporal differences between an uncoded Reference (REF) and a coded Signal Under Test (SUT), aiming at the assessment of small impairments. The basic principle of PEAQ is depicted in Figure 3.17.

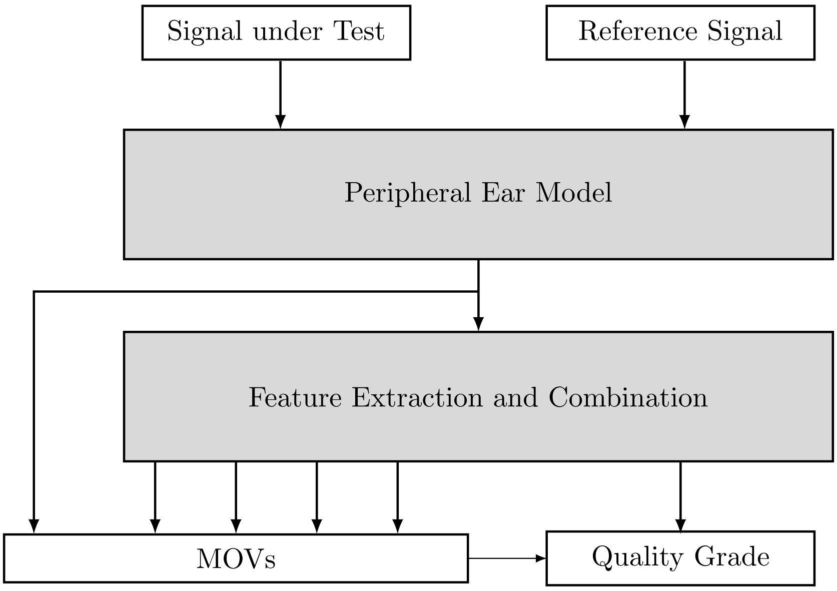


Figure 3.17: High-level representation of the quality prediction stages included in PEAQ (after [276]).

The processing in PEAQ uses the following two stages: The first stage comprises a peripheral ear model that consists of an outer and a middle ear representation extracting various features from the processed audio signals. Based on the excitation patterns as well as the specific loudness and modulation patterns, the Model Output Variables (MOVs) are calculated. Optionally, the second stage employs a subsequent artificial neural network that is intended to emulate higher-level cognitive processes for quality grading, combining the derived MOVs with quality ratings from a human assessor.

PEAQ was later extended for multichannel applications as PEAQ-MC, additionally taking spatial features into account, namely ITD, ILD, and IACC [155]. Being a proprietary extension of PEAQ, PEAQ-MC has not been standardized. In the experiments throughout this thesis, only the auditory models of PEAQ and PEAQ-MC are used. The subsequent neural network for higher-level quality estimation is discarded.

The perceptual model in PEAQ contains all processing stages to transform incoming audio into a basilar membrane representation, the so-called excitation pattern. For this purpose, the peripheral model offers a *basic* and an *advanced* version, where the first uses a Discrete Fourier Transform (DFT) for time-frequency transformation while, additionally, a variable filter bank is employed in the latter. The advanced version is more accurate than the basic version, but computationally less efficient. The processing steps for both basic and advanced version are briefly described in the following.

Basic version: The input signals are transformed into their spectral components using a short-term Discrete Fourier Transform (DFT) with 2048 samples. A Hann-window is moved over the signal with a 50 % overlap between successive frames yielding in a spectral resolution of 1024 lines in the time domain (21 ms @ 48000 Hz sampling rate). Then, the signal is rectified, transforming the complex spectrum into amplitudes over frequency. A specific listening level is simulated by input signal scaling, subsequently adding the frequency response, with f in kHz, of the outer and middle ear using a weighting function. The spectral lines are transformed to power values and grouped into perceptual bands, mapped on a Bark scale with a bandwidth of 0.25 Bark. Then, the internal noise is modeled. Both noise and weighting function model the absolute threshold. The energy distribution is carried out in the frequency domain in two consecutive steps: first, the energy of every frequency band is smeared over the pitch scale; second, the resulting patterns are superimposed. Finally, the time domain spreading is realized by modeling forward masking based on an Infinite Impulse Response (IIR) filter.

Advanced version: Prior to the filter bank, the signal is scaled like in the basic version. A Direct Current (DC) rejection filter is applied, consisting of a 4th order Butterworth high pass filter with a cutoff frequency at $f_{co} = 20\,\text{Hz}$. The advanced version uses a filter-bank-based model consisting of 40 linear-phase filter pairs, one for the real part and the other for the imaginary part. Center frequencies and bandwidths correspond to the properties of the auditory filters. Adjacent filters overlap at the $-6\,\text{dB}$ slopes, defining a bandwidth of 0.6 Bark. A weighted summation between the auditory filter bands is realized as the convolution with a spreading function in the DFT-based model, yielding the frequency domain spreading. A subsequent rectification is applied, calculating the Hilbert transform of the filter outputs and the instantaneous energy. The time-domain spreading, i.e., simultaneous masking, is realized using low-pass filters comprising a raised cosine FIR filter and a 1st-order IIR low pass filter. The first accounts for the ascending, and the second for the descending slopes, hence, representing back and forward masking, respectively. The outputs of both models are the basis for the MOV calculation which is described in more detail in the following paragraphs, after [276].

Model Output Variables: The previously described models calculate 39 MOVs: Twenty-three are derived from the basic, and 16 from the advanced version, including 12 spatial MOVs each.

- **Envelope modulation:** For each filter channel, the temporal envelopes are derived by measuring the modulation from the temporal derivative of the signal envelope. The MOV *avgModDiff1* represents the average (envelope) modulation differences.
- **Modulation difference:** Measuring the difference of changes in the temporal envelopes of REF and SUT, the MOVs *avgModDiff2* and *rmsModDiff* represent the average and Root Mean Squared (RMS) modulation differences, respectively.
- **Partial noise loudness:** In order to attribute a distortion, the respective loudness is measured from the excitation patterns after pattern adaptation. If distortions are missed, then the so-called partial loudness of missing components is derived. It is added to the partial loudness with half of the weight. In the following, the partial noise loudness is described by the RMS noise loudness (*rmsNoiseLoud*).
- **Audible linear distortion:** The measure for linear distortions is analogue to the algorithm used for partial noise loudness. The algorithm is applied before and after the pattern adaptation. The excitation pattern before adaptation is substituted for SUT while REF is substituted by the respective pattern after the adaptation. The average of the linear distortions is represented with the variable *avgLinDist*.
- **Noise-to-mask ratio:** The NMR is derived for every frequency band and is defined as the ratio between error energy and masking threshold [43]. Blockwise, the error signal is derived in the frequency domain for every analysis band by mapping the absolute difference of REF and SUT. If this difference exceeds a certain threshold then the masking flag is set, indicating a possibly audible distortion in this block. The linear average represents the local NMR of a single frame. Based on the local NMR, three features are calculated: the total NMR as the arithmetic mean of the local NMR (*totalNMR*), the segmental NMR as the geometric mean of the local NMR (*segmentalNMR*), and the percentage of distorted frames, if the NMR of at least a single band exceeds a threshold of 1.5 dB (*relDistFrames*).

- **Error harmonic structure:** The autocorrelation of the error energy is calculated based on cepstrum-like analysis. The harmonic structure magnitude is defined as the largest peak of the autocorrelation, and the result is averaged over successive frames. The error of the harmonic structure is represented by *harmStrucOfErr*.

- **Signal bandwidth:** A rough estimate of the signal bandwidths for REF and SUT is calculated for every frame by obtaining the spectral maximum in the range from 21.5 Hz to 24 kHz. The estimated bandwidth is defined by the last frequency line within a frame exceeding the noise floor in energy by 10 dB. Then, the mean over all frames is calculated. The resulting bandwidth estimates for REF and SUT are described by MOVs *bandwidthRef* and *bandwidthSut*, respectively.

- **Detection propability:** The MOV *mfpd* describes the probability of detection of differences between REF and SUT. It depends on the absolute level of the signals, where the signal with higher level defines the reference level. Here, for each frame and each band, the JNLDs are estimated. If the level difference is equal to the JNLD, then the detection probability is set to 0.5. The probabilities of all bands are combined in each frame to obtain the detection probability for the respective frame, whereas the total probability is obtained by finding the maximum of the smoothed local detection probabilities.

- **Average distorted block (or frame):** This MOV is related to the detection probability and describes the number of frames that are probably distorted, i.e., if the probability of detection exceeds a certain threshold for that frame. The MOV is represented by the variable *avgDistBlk*.

- **Spatial features:** In addition to the MOVs of the basic and advanced version, also binaural cures are evaluated based on PEAC-MC [155]. Namely, ITD, ILD, and IACC are calculated for every time frame and frequency, resulting in four MOVs each. All frequencies are evaluated for every time slot using the average and the maximum frequency value. This is done for the average and maximum over all time slots, resulting in four MOVs: namely, *meanFmaxT*, *maxFmaxT*, *meanFmeanT*, and *maxFmeanT* are used in the evaluations.

Table 3.2 provides an overview of all 39 MOVs used in the experiments throughout this thesis.

Table 3.2: MOVs of PEAQ model with respective abbreviation.

#	MOV	Abbreviation	Version
V1	winModDiff1	Modulation Differences	basic
V2	avgModDiff1	Average Modulation Differences	basic
V3	avgModDiff2	Average Modulation Differences	basic
V4	rmsNoiseLoud	Noise Loudness (RMS)	basic
V5	BandwidthRef	Bandwidth REF	basic
V6	BandwidthSut	Bandwidht SUT	basic
V7	totalNmr	Total NMR	basic
V8	relDistFrames	Relative Distance between Frames	basic
V9	mfpd	Detection Probability	basic
V10	avgDistBlk	Average Distorted Block	basic
V11	harmStrucOfErr	Harmonic Structure of Error	basic
V12	rmsModDiff	Modulation Differences	advanced
V13	rmsNoiseLoudAsym	Noise Loudness (RMS)	advanced
V14	avgLinDist	Average Linear Distortions	advanced
V15	segmentalNmr	Segmental NMR	advanced
V16	ildMaxFMeanT$_{adv}$	Interaural Level Differences	advanced
V17	ildMaxFMaxT$_{adv}$	Interaural Level Differences	advanced
V18	ildMeanFMeanT$_{adv}$	Interaural Level Differences	advanced
V19	ildMeanFMaxT$_{adv}$	Interaural Level Differences	advanced
V20	itdMaxFMeanT$_{adv}$	Interaural Time Differences	advanced
V21	itdMaxFMaxT$_{adv}$	Interaural Time Differences	advanced
V22	itdMeanFMeanT$_{adv}$	Interaural Time Differences	advanced
V23	itdMeanFMaxT$_{adv}$	Interaural Time Differences	advanced
V24	iaccMaxFMeanT$_{adv}$	Interaural Cross Correlation	advanced
V25	iaccMaxFMaxT$_{adv}$	Interaural Cross Correlation	advanced
V26	iaccMeanFMeanT$_{adv}$	Interaural Cross Correlation	advanced
V27	iaccMeanFMaxT$_{adv}$	Interaural Cross Correlation	advanced
V28	ildMaxFMeanT$_{bas}$	Interaural Level Differences	basic
V29	ildMaxFMaxT$_{bas}$	Interaural Level Differences	basic
V30	ildMeanFMeanT$_{bas}$	Interaural Level Differences	basic
V31	ildMeanFMaxT$_{bas}$	Interaural Level Differences	basic
V32	itdMaxFMeanT$_{bas}$	Interaural Time Differences	basic

V33	itdMaxFMaxT$_{bas}$	Interaural Time Differences	basic
V34	itdMeanFMeanT$_{bas}$	Interaural Time Differences	basic
V35	itdMeanFMaxT$_{bas}$	Interaural Time Differences	basic
V36	iaccMaxFMeanT$_{bas}$	Interaural Cross Correlation	basic
V37	iaccMaxFMaxT$_{bas}$	Interaural Cross Correlation	basic
V38	iaccMeanFMeanT$_{bas}$	Interaural Cross Correlation	basic
V39	iaccMeanFMaxT$_{bas}$	Interaural Cross Correlation	basic

3.3 Predictive Modeling

According to [141], predictive modeling is defined as “the process of developing a mathematical tool or model that generates an accurate prediction”. It is commonly used for regression or classification tasks, the former predicting a continuous-valued attribute of an object and the latter identifying the category an object belongs to. While most models can be applied for both purposes, their preparation and evaluation differ between regression and classification, as will be shown later. Also, each task is subject to specific problems. In practice, the right model choice depends on the nature of the prediction problem and is always a trade-off between prediction accuracy, model interpretability, and its computational complexity.

Basically, predictive modeling can be divided into two major categories: supervised and unsupervised prediction. In the first, the prediction is based on attributes describing the data while in the latter, the data comes with no additional information at all. Supervised prediction is commonly used for classification and regression problems, whereas unsupervised prediction is applied to discover groups or determine data distributions, also referred to as *clustering* or *density estimation*, respectively. All models employed in the experiments throughout this thesis are briefly described in the following. For further reading on predictive modeling, refer to, e.g., [33, 117, 141, 173].

3.3.1 Common Predictive Modeling Techniques

The respective literature presents a variety of predictive models. However, in the scope of this thesis, only a selection of commonly used robust and powerful models are applied. The following overview is based on [141], describing the most prominent linear, nonlinear, and tree-based models for classification and regression analyses.

Linear modeling techniques

Linear modeling techniques show a linear behavior in their parameters, which can mathematically be described by [141]

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + ... + +b_p x_{ip} + e_i, \tag{3.75}$$

where y_i is the response of the i-th sample, b_j represents the estimated coefficient for the j-th predictor, x_{ij} the value of the j-th predictor for the i-th sample, and e_i the random error which cannot be explained by the model. The advantages of linear (regression) models are their low complexity and easy interpretability. However, their estimates are influenced by outliers, i.e., samples far away from the overall data trend, which can significantly lower their predictive performance. Although such models show a linear behavior in their parameters, it is also possible to adapt them to nonlinear relationships, but only if the nonlinearity in the data is known [141]. The accuracy of linear models is further limited if multiple predictors show nonlinear relationships.

The following paragraphs describe common linear models, comprising simple models based on linear and partial least squares regression. Their counterparts for classification are known as linear and partial least squares discriminant analysis. Although further linear methods for regression and classification applications are given in [141] such as penalized models such as ridge regression and lasso, or models based on nearest shrunken centroids, they are not further addressed in the presented analyses.

Linear regression: Linear Models (LM) are the simplest form of prediction models comprising ordinary linear regression, Robust Regression (RLM) or Generalized Regression (GLM). All seek for functions best approximating model parameters in a way that e_i is minimized. An ordinary LM tries to fit a plane in the data with the aim to minimize the Sum-of-Squared Errors (SSE) between observed and predicted response. If somehow a nonlinear relation between the predictors is present then more robust versions of LMs are applicable such as RLMs or GLMs. If outliers corrupt the regression accuracy, an approach to improve the model robustness is to use an alternative metric to SSE for residual reduction. For example, the Huber function uses the squared value for small residuals below a specified threshold and the simple prediction-observation difference above that threshold, consequently reducing the influence of outliers. RLM, on the other hand, relies on so-called M-estimators, i.e., maximum-likelihood estimators, for residual reduction [228, 282]. These linear models also behave differently depending on the data distribution. While general LMs relate to normally distributed data, the distribution of the response variable in GLMs can be a member of the exponential family, like binomial, Poisson or other similar distributions [189]. GLMs comprise two additional components, the so-called link and variance functions which relate the model to the response variable. Respectively, they describe how the mean depends on the linear predictor and how the variance depends on the mean.

In classification applications, models for Linear Discriminant Analysis (LDA) try to minimize the probability of misclassification by evaluating class probabilities and multi-variate distributions in the data. LDAs determine the respective discriminant boundaries based on linear combinations of the predictors such that the between-group variance is maximized compared to the within-group variance [141], like using the Karhunen–Loève transform [131]. Mathematically, this is achieved by relating their covariance matrices, which results in an optimization problem that can be solved with an eigenvector corresponding to the largest eigenvalue. This vector represents the linear discriminant based on which a new sample is categorized. The amount of discriminant vectors is the tuning parameter to find an optimally performing model. In practice, LDA classification performance is negatively influenced by near-zero-variance predictors and collinearity. Respective data preprocessing can increase the LDA performance, like using a Principle Component Analysis (PCA)[6] [122, 131] for dimensionality reduction.

Partial least squares: Partial Least Squares (PLS) regression is based on the Nonlinear Iterative Partial Least Squares (NIPALS) technique which linearizes models that are nonlinear in their parameters [300]. PLS seeks for linear combinations between predictors in a way that maximally summarizes the covariance with the response and, in addition, requires the resulting components to have maximum correlation with the outcome [141]. This approach makes PLS a robust solution when predictors are highly correlated or their amount exceeds the observations. The number of components used for modeling is the respective tuning parameter. However, as in case of PCA, the data should be centered and scaled beforehand. Rather than using PLS for regression, it is also applicable in classification tasks where it is referred to as PLS discriminant analysis. It is commonly applied when LDA breaks, for example, due to collinearity. However, instead of using a PCA like in LDA, the discrimination between classes is based on PLS.

[6] Note that the PCA is similar to the Karhunen–Loève transform when dealing with discrete data. Likewise, the PCA transforms a large set of partly interrelated variables into a new set of uncorrelated variables, the so-called Principle Components (PCs), with the first PC accounting for the highest variances, the second for the second highest variances and so forth.

Nonlinear modeling techniques

Nonlinear models should be used if nonlinear relationships are present in the data, which can not be accounted for with linear models. In such cases, nonlinear models usually achieve high prediction performances, but, on the downside, are more complex than linear models and consequently provide less computational efficiency and interpretability.

Nonlinear discriminant analysis: Nonlinear Discriminant Analysis (NDA) comprises classification methods to separate the data based on nonlinear structures such as the Quadratic or Regularized Discriminant Analysis methods, QDA and RDA [93], respectively. They are extensions of NDA, with QDA relying on quadratic surfaces to separate the data whereas NDA uses hyperplanes. RDA is used when the data is best split with a surface that lies between a linear and quadratic one. Also, a mixture of these methods can be used for classification tasks, the Mixture Discriminant Analysis (MDA) [116].

Artificial neural networks: Common nonlinear prediction models applied for regression and classification tasks are Artificial Neural Networks (ANN) [32, 229, 277]. Figure 3.18 shows a schematic of an ANN with its distinct layers: an input layer, hidden layer, and outpupt layer.

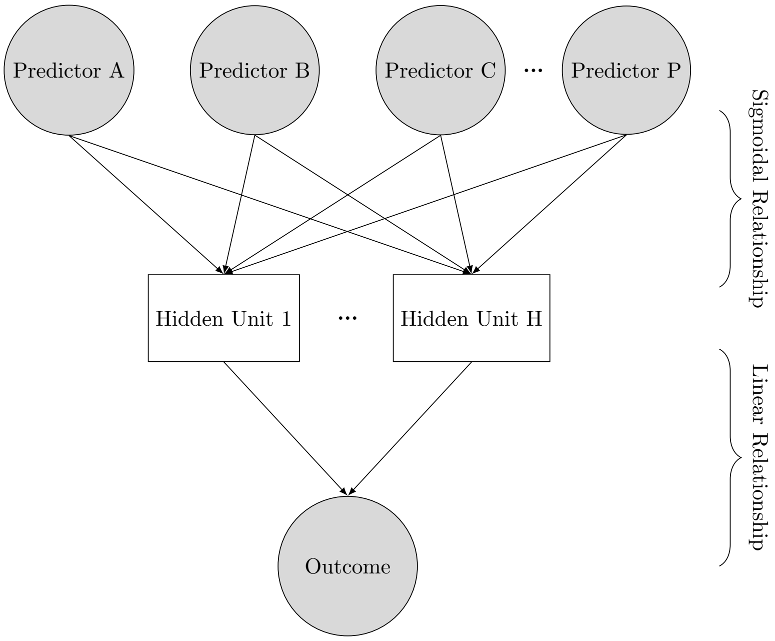


Figure 3.18: Schematic of an artificial neural network with its input layer, hidden layer, and outpupt layer (after [141]).

ANNs are inspired by information processing in the human brain. Connected by synapses, the neurons are arranged in layers; an input layer comprising all covariates as separate neurons, i.e., P predictors, a hidden layer which is unknown, including H hidden units, and an output layer providing the response variables [32]. The complexity of the model is determined by the number of hidden layers. Known as *activations*, linear combinations of the input variables are weighted and biased in the first layer. Each of these activations is subsequently transformed with a differentiable nonlinear function which results in hidden units. In practice, the nonlinear functions are generally sigmoidal or tanh functions from which the output is again linearly combined, yielding in the output parameters of the network. Such a network is called a *feed-forward* ANN since the information propagates unidirectionally through the network. In order to improve the performance, ANNs are trained using backpropagation algorithms which is a standard algorithm when training ANNs today. They compare the predicted to a given output and adjust their parameter weights accordingly, in order to find a local minimum of an error function, as in case of the Sums of Squares Error (SSE).

ANNs for classification are similar to ANNs for regression regarding their basic structure and components. However, instead of a single output, the bottom layer of ANNs for classification comprises multiple output nodes, with one for each class. In addition, a nonlinear transformation is used on the hidden units. Each class prediction is based on a linear combination of the hidden units, which have been transformed by a sigmoidal function to values between 0 and 1. Note that these values are no probability values as they don't add up to 1. Applying a *softmax* function transforms these outputs into probability-like values. Like their regression counterparts, ANNs for classification tasks are often subject to over-fitting (see Section 3.3.6). This effect, however, can be reduced by means of model-averaging, an approach to stabilize the prediction by creating several models with different starting values and averaging the results. In addition, collinearity and non-informative predictors affect the classification performance of ANNs.

Note that the experiments conducted in this thesis mainly rely on *classic* neural networks. More advanced approaches, like deep learning using Convolutional Neural Networks (CNN), are not applied, but proposed for future work.

Multivariate adaptive regression splines: Multivariate Adaptive Regression Splines (MARS) is a nonparametric regression method that makes no assumptions on the underlying predictor-response relationships [91]. It uses surrogate variables as predictors which are functions of one or two predictors at a time. These features basically break the predictor into two groups, modeling a linear relation between the predictor and the outcome in each group [141]. These two new features are so-called *hinge* functions, modeling slopes and intercepts by ordinary linear regression in a way that each feature models a portion of the original data. The cut point is determined for each data point and each predictor, a linear regression model applied with the candidate features, and the corresponding modeling error calculated. The cut point with the least error is then used for modeling. This process is iteratively repeated for each feature set until a user-defined stopping point is reached. Finally, a pruning procedure evaluates every predictor in terms of its contribution to the model performance, subsequently removing predictors with low influence. This procedure was based only on single predictors. Increasing the model-degree, MARS is also able to build features including multiple predictors. The degree and the number of model terms are the tuning parameters in the model building process. The described technique makes MARS a powerful modeling tool as it inherently conducts feature selection in a way that the used features are directly connected to the model performance. Also, MARS models are easily interpretable due to their use of linear modeling in the predictor space. Furthermore, these models are robust against correlated or near-zero-variance predictors. MARS shows good prediction results for moderate sample sizes ($50 \leq N \leq 1000$) and dimensions ($3 \leq n \leq 20$) [92]. Flexible Discriminant Analysis (FDA) is a technique to extend MARS to perform classification tasks. This is achieved by fitting a set of regression models to binary class indicators which result in discriminant coefficients.

Support vector machines: Support Vector Machines (SVM) [280] are highly flexible nonlinear regression models which were originally designed for classification tasks. However, they are also applicable for robust regression analyses [281]. The original problem in support vector classification is the separation of two classes based on a set of training samples. Both classes, i.e., both sets of vectors, are optimally separated by a hyperplane when the error is minimal and the distance of the closest vector to the hyperplane is maximal. This distance is called *margin* and the closest data points are the *support vectors*, hence the name support vector machine. Consequently, SVMs only use a subset of samples from the training data for modeling. The optimal separating hyperplane is a linear classifier which is based on a user-defined threshold ϵ. This threshold is accounted for in the ϵ-loss function which is the cost tuning parameter in the SVM model building process. If samples fall within this threshold then they are not considered in the fit, whereas samples outside are weighted on a linear scale. Therefore, SVMs are robust against outliers as samples with large residuals only have a limited effect on the regression equation, whereas samples with small residuals have no effect at all. SVMs become very flexible if the cost-parameter is large, but also get more sensitive to errors which might lead to over-fitting. Small values stiffen the model, making it prone to under-fit the data. However, centering and scaling the input data is suggested in [141] for better performance. Depending on the nature of the prediction problem, SVMs can use different kernel functions to split the data, like polynomial, radial-based, or certain sigmoid functions. This so-called *kernel-trick* maps the input data into a (possibly very) high-dimensional feature space where an optimal separating hyperplane[7] is constructed applying the standard linear SVM regression method [109]. The corresponding back-transformation into the lower dimensional output space reduces the data to a small number of support vectors.

K-nearest neighbor: K-Nearest Neighbor (KNN) regression and classification is based on the K-closest samples from the training set when predicting a new sample [7]. The predicted response for the new sample can be the mean or median K-neighbor's response. The method to calculate the distance between samples is user defined and can be, for example, the Euclidean, Minkowski or Manhattan distances. Since KNNs measure distances between samples, centering and scaling (see Section 3.3.3) the data should be applied beforehand. An important tuning parameter is K, the number of neighbors. For small values of K, KNNs tend to over-fit the data while large values might lead to under-fitting. In the context of classification, some unique aspects have to be highlighted for KNNs. They predict the class of a sample evaluating its neighbors classification, which is estimated based on the Euclidean or Minkowski distance. The class of a new sample is determined by the highest probability estimate which is calculated by the proportion of neighbors in each class. In practice, a major drawback using KNNs is their high computational complexity and therefore, limited real-time applicability.

Decision trees and rule-based models

Decision trees were also identified as powerful tools for regression and classification with advantageous properties [46]. They are highly interpretable and easy-to-implement tools for regression and classification. Decision trees can handle various predictor types which may be continuous, categorical, etc. In addition, they are robust preduction models against missing, sparse, or skewed data. Therefore, no data preprocessing is

[7] A hyperplane is a subspace of dimension $n - 1$ in n-dimensional ambient space, i.e., a hyperplane is a 2D-plane in 3D-space or a line in 2D-space, respectively.

required. Trees also do their own feature selection. If a predictor is never used in a split, i.e., the decision what branch should be taken, then it is discarded from the tree building process. On the downside, trees often lack stability which might be introduced by highly correlated predictors, resulting in poor prediction accuracies. To account for this weakness, ensembles, bagged trees, or boosted trees were introduced [141].

Basic regression trees: Trees are cascades of one or more *if-then*-statements based on which the data is partitioned. Each so-called split separates the data into terminal nodes or leaves [46]. For new samples, the statements defined by a tree are followed down to a terminal node according to the respective predictor values. The resulting structure defines the prediction model. The basic regression tree building process starts with a split of the entire data set, partitioning the data into two groups. Based on the minimum of the overall SSE, the best split is searched for over all predictor values. This process is recursively repeated for all sub-groups until the amount of samples left in a split reaches a certain threshold. To strengthen a grown tree against over-fitting, so-called pruning (shrinking) can be applied to reduce the complexity of the model and increase its interpretability. Multiple, possibly interrelated statements are called a *rule* and can further increase the accuracy of the tree. To reduce the complexity of a tree and improve its performance, pruning can reduce the tree depth by combining multiple rules into a single one. A number of different rule-definition strategies are proposed in the literature.

Ensembles: Ensembles are methods combining many model predictions into a single one in order to build a more stable model and reduce the effect of over-fitting [45]. Bagging, for example, uses bootstrap aggregation (see the resampling methods in Section 3.3.4) to construct an ensemble. So-called random forests were introduced to further improve the performance of bagged trees by adding randomness for decorrelation between trees. This is achieved by random split selections, random descriptor subsets or simply by adding noise. Bagged trees, however, are computationally more complex and less interpretable than their basic versions.

Boosted trees: Enhanced versions of ensembles are called boosted trees. They combine multiple weak classifiers into a strong one. This so-called *boosting* is based on loss- or cost-functions which are related to the tree performance via model tuning parameters. Boosting seeks to minimize these functions in order to provide an optimal parameter set for the given modeling task. Common boosting techniques are AdaBoost (ADA) [90], Stochastic Gradient Boost (SGB) [94], and C5.0 boosting (C50) [141].

3.3.2 Model Performance Evaluation

Depending on the modeling problem, several methods are given in the literature to evaluate the prediction performance. These methods are fundamental for feature selection, model tuning, and the predictor importance evaluations, i.e., how strongly a predictor contributes to the model decision. The descriptions follow [141].

Regression performance

In regression tasks, commonly three measures are used for objective assessment. First, the prediction accuracy is evaluated based on the correlation of the original and predicted PDG mean, i.e., the correlation coefficient R. Another common performance measure is the coefficient of determination R^2 which is the squared correlation coefficient. It measures the amount of information explained by the model. In case of ranking problems, the model performance is usually evaluated based on Spearman's rank correlation, evaluating the correlations between observed and predicted ranks. An additional metric is the Root Mean Square Error (RMSE) which is calculated by

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})} \ , \tag{3.76}$$

where n is the number of samples, y_i the real, and $\hat{y_i}$ the predicted one. It is a function of the model residuals, measuring the distance between observed and predicted value.

Note that R^2 is a measure for correlation which can only serve as a hint on prediction accuracy. Therefore, it is recommended to also visualize the results and observe original and predicted data.

Classification performance

Compared to regression, model performances are evaluated differently in the context of classification, since such models generate continuous class probabilities and discrete class categories. The first describes the model's confidence about the classification and the second categorizes this class. Other models, for example like ANNs, produce continuous values which do not sum up to 1 like class probabilities do. They are transformed into probability-like values using the *softmax transformation* after [47], representing the respective class probabilities. Classification trees, however, are evaluated differently. Their accuracy is measured in purity of a node with respect to one class. The *Gini*-Index, for example, takes class probabilities into account and is a suitable measure for purity.

The most straightforward approach to evaluate the model performance is Overall Accuracy (OA), measuring the agreement between observation and prediction. However, the problem of class imbalances (discussed later) cannot be accounted for with this measure. The (weighted) *Kappa*-Statistic is an approach to achieve this, as it also takes the sample frequency within a class into account [61]. In addition, the No-Information Rate (NIR) is a suitable measure to identify class imbalances. It defines the prediction accuracy which can be achieved without a model and can alternatively be understood as the percentage of the largest class in the training set. If OA is higher than NIR, then the model performance might be reasonable. A more detailed information on the classification performance can be given with a confusion matrix. It is a simple table, cross-listing the observed and predicted classes where the diagonal cells show correct predictions while all other cells denote the respective erroneous classifications. The entries in all non-diagonal cells are used to describe the Sensitivity (SEN) and Specificity (SPEC) of the model. SEN is often referred to as the *true positive rate*, indicating that an event is correctly identified for all samples from this event. Conversely, SPEC is the rate that non-event samples are correctly categorized as non-events. The Receiver Operating Characteristic (ROC) curves represent the trade-off between SEN and SPEC, combining both into a single value represented by the Area Under the Curve (AUC) which indicates the model performance.

3.3.3 Data Preprocessing

To improve the prediction performance of the model, the data to be analyzed should be preprocessed thoroughly. Common approaches for feature selection and methods for data preprocessing are briefly discussed in the following, after [141].

Predictor selection: A careful predictor choice can significantly improve the model performance. Appropriate measures and guidelines for selection can be simple between-predictor-correlations or evaluating (multi-)collinearity between predictors with, for example, a PCA.

In the literature [141], different methods exist for automatic predictor evaluation and selection: the so-called *filter* method and the *wrapper* method. The first evaluates the relevance of predictors fulfilling a certain criterion, like a plausible relationship to the outcome of the model, while the latter comprises predictor search algorithms, testing different predictor combinations by adding or removing predictors and comparing the resulting accuracy. The search is commonly based on forward, backward or stepwise selection. The metric used to decide whether a predictor is chosen or not can vary, like p-values indicating significance of a predictor, or Akaike Information Criterion (AIC) [5], which is a penalized version of the SSE.

Removing skewness: Distributional skewness describes the asymmetry of data points, which can affect the model performance in practical applications. It is assessed with the skewness statistic which reads [141]

$$\text{Skewness} = \frac{\sum(x_i - \bar{x})^3}{(n-1)v^{3/3}}. \tag{3.77}$$

In the equation, x is the predictor variable, $\bar{x}$ the sample mean and n the number of values. Symmetric distributions provide near-zero skewness values, while right skewness is indicated by positive, and left skewness by negative values. Skewed data can be problematic for some models, so the application of appropriate data transformations for skewness removal is suggested [141]. Such transformations are, for example, the log-function, square root or the inverse. However, the so-called Box-Cox-transformation after [37] is recommended as a straightforward approach for skewness removal as it is robust against numerical issues [141].

Data centering and scaling: Centering and scaling (C&S) is important to identify the underlying relationships within the data without bias from their scales [141]. This improves the numerical stability of a model but, on the downside, reduces the interpretability of the data as their original units get lost. Predictor variables are centered by subtracting the mean from all variables, and scaled by dividing each variable value with its standard deviation. C&S is recommended, for example, prior to performing a PCA.

3.3.4 Model Training and Tuning

In order to train and tune a model, predictive modeling is commonly based on the following procedure: to build a model, the data is divided into two subsets: a training set and a test set. The first is used to train the model on the data characteristics and to tune it for optimal performance while the second set is held out to see how the trained model actually performs when predicting unknown samples. The test data serves for model testing. Further information on data splitting strategies are described in [141]. Model tuning is commonly based on resampling, a technique to increase the variance in the data during model training and thereby reduce the chance of over-fitting.

Resampling techniques

Resampling[8] is used to generalize the prediction performance of the model, estimating its accuracy during training. In practice, these techniques randomly split the data into a subset for training and a test set for evaluation, subsequently evaluating the model performance against the test set. This process is repeated multiple times, the results aggregated and summarized to provide a generalized estimate for model performance.

Although further methods exist, the following resampling techniques are commonly chosen in practical applications [141]:

- **k-fold cross-validation:** The data is allocated into k subsets of equal size. The model is fit to the first subset and evaluated predicting the remaining samples. This process is repeated for each subset. In practice, k is usually set to 5 or 10, whereas it was stated in [141] that $k = 10$ provides a lower estimation bias. Repeating the k-fold cross-validation n-times can further increase the precision. An alternative version, for example, is the leave-one-out cross-validation where k equals the number of samples.

- **Repeated training/test splits:** The so-called *Monte Carlo cross-validation* creates multiple training and test sets for an arbitrary number of repetitions. For a stable performance, 50–200 repetitions are suggested in the literature [141].

- **Bootstrapping:** In bootstrapping, the test set is removed from the original set and the training samples are randomly chosen from the remaining data in a way that the training set has the same sample size as the original set. This means that the training set contains multiple instances of the same samples.

Although performance estimates are likely to vary for each resampling method and each model, the 10-fold cross validation is recommended in [141] as the resampling method of choice for practical applications. In most cases, it provides similar or even better performance estimates than other resampling methods. Additionally, it is the most efficient method in terms of computational efficiency. Therefore, it is applied for resampling in the experiments throughout this thesis.

Based on resampling, the model is tuned by adjusting its tuning parameters to achieve the best prediction accuracy. Model tuning is described next.

[8] Note that, in the context of this thesis, resampling means statistical resampling and should not be confused with, for example, resampling in signal processing.

Model tuning

Model tuning describes the process to find the best set of (tuning) parameters for a model to achieve best prediction performance. For tuning, the model is trained with the training data and tested against the hold-out set based on resampling. For each resampling split, different tuning parameters are evaluated in terms of their achieved accuracy, based on which the best fitting model is identified. Which tuning parameter is optimal for the modeling task can be decided on different metrics. In classification applications, this can be the estimated prediction performance, i.e., the overall accuracy or the *Kappa*-statistic, while in regression tasks, tuning is mainly based on R^2 or RMSE measures. For some models, these parameters are also referred to as the *complexity* or *cost parameters*. They describe the complexity of the model, like for example the size of a forest or the amount of hidden units in ANNs. In this case, higher cost values result in lower computational efficiency.

3.3.5 Predictor Importance

In practice, it is often useful to know which predictors are relevant for the given prediction task, and how these predictors contribute to the model performance. This can be achieved in a model-dependent way, since most models provide built-in measures for predictor importance. MARS and decision trees, for example, monitor their performance while adding or removing predictors. Other methods such as linear or logistic models use model coefficients or statistical measures to choose predictors, like the t-statistics. The relationships between predictor and outcome can be evaluated. This, however, provides no information on the predictor-model relationship.

For prediction models that generate numerical outcomes, the classic approach is based on simple correlation or rank correlation metrics in order to quantify the relationship with the outcome. If the relationships are nearly linear or curvilinear, then the Spearman's correlation coefficient is more effective. For nonlinear relationships, more flexible methods are recommended, like the Locally-weighted Regression Model (LOESS). Its modeling approach is based on a series of polynomial regressions resulting in a LOESS pseudo-R^2 statistic [59]. These measures only evaluate single predictors, so they can not identify groups of interrelated predictors. In classification tasks where the outcome is categorical, different techniques are applicable: For two-class problems, the area under the ROC (AUC) can give information on predictor importance. AUC would be 1 if a predictor can perfectly separate the two classes, and would be approximately 0.5 if it is completely irrelevant. The *one versus all*-ROC curve extends ROC for multi-class problems [113, 206, 104]. Here, the AUC is evaluated for every class, providing the overall relevance of a predictor based on the average (or maximum) of all AUCs. Further metrics are, for example, the *Relief* method [134] or its extension, the *ReliefF* method [138], both being able to recognize nonlinear relationships between predictors and outcome. The so-called Maximal Information Coefficient (MIC) is another common measure describing the relationship between two variables [227].

3.3.6 Problems in Predictive Modeling

In practice, several pitfalls—inherent to predictive modeling—can impair the performance of the model [141]. Such pitfalls can be, for example, outliers in the data or unjustified extrapolation. The latter occurs when prediction models are developed and trained for a specific purpose, i.e., based on a certain type of data from a particular population; applying the same model to data from a different population is likely to result in inaccurate predictions. In addition, measurement errors in the data, both for predictors and observations, can also decrease the performance of the model. Therefore, apart from a thorough data basis, care has to be taken in the model design process which is especially critical if only a small (or incomplete) amount of data is available. The most prominent problems are briefly described in the following [141].

Inadequate data preprocessing and model validation: Depending on the applied model type, preprocessing of the data can have a significant impact on the model performance. A careful selection of predictors is important for optimal model performance and can for example be achieved by adding relevant, or discarding irrelevant predictors which provide no or redundant information. Interrelationships between two predictor variables are called *collinearity*, and *multicollinearity* when more than two variables are related. Removing redundant predictors reduces complexity and leads to a more interpretable model without reducing the prediction performance.

Over-fitting: A common error in predictive models is over-fitting. It means that the model has been fitted too well on the training data, which often occurs when the model is trained with small sample sizes. In such cases, the model too accurately accounts for the noise in the data instead of the underlying relationship. Over-fitted models usually show poor accuracy when estimating new, unknown samples. Resampling can decrease the chance of over-fitting and provide a more realistic estimate on model performance. It also allows to generalize how the model would perform on new samples.

Models are specifically prone to over-fitting if only a small amount of samples is available for modeling. To account for this weakness, resampling can significantly increase model performance, like the 10-fold cross-validation which is explicitly recommended in [141] if the sample size is small. A repeated cross-validation can further improve the results for generalization.

The curse of dimensionality: Another common problem is the so-called *curse of dimensionality* [19]. It refers to the fact that increasing the number of predictor variables leads to an improvement of the model only up to a specific point. Beyond that point, the model performance decreases rapidly when adding more predictors [32]. This is due to the fact that the amount of data required for accurate generalization grows exponentially as the amount of dimensions increases. The direct result will be an over-fitted model. For example, a classifier trained with too much features will most likely learn exceptions that are specific to the training data and will not generalize well. Appropriate data preprocessing is therefore crucial to reduce the risk of the curse of dimensionality.

The variance-bias trade-off: A problem specific to regression models is the so-called variance-bias trade-off. It states that simple models show low variance, they under-fit. This means that the model would not substantially change when adapting to new samples, but, as a consequence, it shows high bias as it is not effective in modeling the new data. More complex models, on the other hand, are subject to increased variance but provide lower bias as they adapt to the pattern of new samples. However, models with very high variance are prone to over-fitting [141].

Class imbalances: In classification tasks, it can occur that one or more classes are over- or underrepresented to others. These so-called class imbalances can affect the prediction performance as the model is likely to fit on the overrepresented class [141]. To strengthen the model against class imbalances, different methods are at hand, like tuning the model to an increased accuracy when predicting the minority class, or to find an optimal balance between sensitivity and specificity. With a priori knowledge of class imbalances, it is recommended to do a balanced data split before model training.

4 Perceptual Evaluation

This chapter presents the perceptual evaluation of spherical microphone array auralizations, addressing various aspects of the measurement system by means of quantitative and qualitative analysis. In particular, three listening experiments are conducted: The first two quantitatively assess the influence of measurement errors such as spatial aliasing, transducer noise, and microphone positioning offsets in free-field and reflective environments while the third investigates different array configurations in a qualitative analysis using ASW and LEV as quality factors. This is done for different sound field orders and varying reflective environments. The resulting perceptual data forms the basis for predictive modeling in Chapter 5. The following experiments are conducted:

- Experiment I evaluates array measurement errors under free-field conditions based on a quantitative analysis.
- Experiment II evaluates array measurement errors in simple room geometries based on a quantitative analysis.
- Experiment III provides a descriptive analysis of various array configurations in simulated sound fields based on ASW and LEV attribution.

In all experiments, the binaural synthesis employs HRTFs of a Neumann KU100 dummy head [27]. An extra-aural headphone with equalized transfer characteristics [80] is used for playback. No head-tracking is applied, and the assessors were briefed to keep their head still during the experiment. The listening test laboratory fulfills the requirements for a listening room as specified in ITU-R BS.1116 [50]. In all three experiments, ten test signals from sound quality assessment material (SQAM) data base [79] are used as audio material. The signals are chosen to cover a variety of realistic audio content which would be presented in concerts or other audio performances, comprising male and female singing voices, popular and orchestral music as well as samples from single instruments such as dry recordings of a clarinet, an organ, or drum sounds. Also critical test material such as castanets is included in the experiments. The following Table 4.1 lists the ten audio signals used in Experiments I, II, and III.

Table 4.1: List of test signals used in listening experiments I, II, and III with respective track number, signal type description, dryness of the recording, and length.

Signal #	Name	Type	Reverberation	Length [s]
1	ctrack27	castanets	reverberant	19
2	ctrack20	clarinette	dry	38
3	ctrack44	female singer	dry	27
4	ctrack58	guitars	dry	15
5	ctrack70	pop music I	reverberant	20
6	ctrack47	male singer	dry	29
7	ctrack64	orchestra	reverberant	30
8	ctrack56	organ	reverberant	33
9	ctrack12	pop music II	reverberant	26
10	ctrack30	toms	dry	25

The following paragraphs briefly describe the corresponding quantitative and descriptive analysis methods used to evaluate the experimental data.

Quantitative analysis method: The evaluation of the quantitative data derived in the presented experiments is based on analysis of variance (ANOVA). ANOVA is a statistical method for parametric data analysis, evaluating variabilities with the aim to describe inequalities among population means [205]. More specifically, ANOVA evaluates if the independent variable, e.g., a system error such as spatial aliasing, has an effect on the dependent variable such as the ratings of an assessor. Based on the variability between different parameters of the independent variable, ANOVA allows statements about significance, i.e., whether the measured effect was caused by that variable or not[9]. If the data is nonparametric, a combination of the Friedman or Wilcoxon tests is proposed as an ANOVA alternative [269].

However, in order to apply an ANOVA test, some general assumptions need to be fulfilled [269]: firstly, normally distributed data per test parameter is required which can be checked using a Kolmogorov-Smirnov or a Shapiro-Wilk Test; secondly, the data must be interval data; and thirdly, homogeneity is required among all parameters (which can be tested with a Levene's Test). If one of these requirements is violated, for example, due to non-normally distributed data, then nonparametric analysis methods have to be applied such as the Friedman Test [269]. For more information on statistical data analysis, the reader is referred to, for example, [267, 290, 205, 66].

[9] ANOVA checks if the null-hypothesis H0 holds, i.e., two or more population means are equal [205], which means that no significant difference is present.

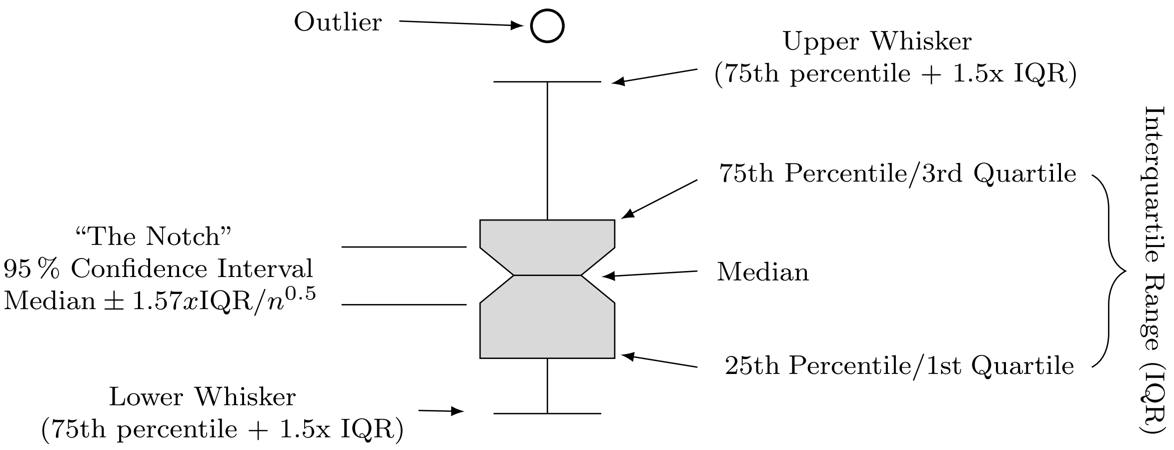

Outlier
Upper Whisker
(75th percentile + 1.5x IQR)
75th Percentile/3rd Quartile
"The Notch"
95 % Confidence Interval
Median ± $1.57x$IQR$/n^{0.5}$
Median
25th Percentile/1st Quartile
Lower Whisker
(75th percentile + 1.5x IQR)
Interquartile Range (IQR)

Figure 4.1: Schematic of a notched box plot and its interpretation (after [56]).

The assessor's ratings, i.e., the perceptual difference grades (PDGs), are presented by notched box plots as shown in Figure 4.1, following [56]. The size of the box shows how confident assessors are in their ratings, with small boxes indicating low variance and therefore high confidence while larger boxes represent lower confidence. The box is limited by upper and lower quartile, the 25th and 75th percentiles, which represent the so-called interquartile range (IQR) that comprises 50 % of the population. The black line within the box depicts the median of the distribution while the notches indicate the 95 % confidence intervals[10]. For normally distributed data, the whiskers extend the variability of the data above and below the 75th and 25th percentile by 1.5 × IQR, respectively, comprising 99.3 % of the data. Outliers are marked by a black circle. Statistical significance is assumed between two conditions if their confidence intervals, i.e., their notches, do not overlap. If this can not clearly be stated, then an additional significance test will be conducted. After testing if the data is normally distributed or not, a significance analysis as described above is carried out for each condition pair. A horizontal square bracket marks the significant difference between two conditions in a plot.

Qualitative analysis method: The qualitative analysis conducted in Experiment III is based on descriptive attributes describing the sensory quality. Their strength is evaluated and quantified on a 100-point continuous scale, also resulting in PDG ratings.

[10] Although evaluating significances by comparing notches is an informal test, the derivation of the notch lengths is based on the formal concept of a hypothesis test. If the data is normally distributed, then the notches provide an approximate 95 % hypothesis test that their medians are equal [56].

4.1 Experiment I: Quantitative Error Analysis under Free-Field Conditions

Experiment I quantitatively analyzes the effect of measurement errors in free-field sound fields. Specifically, spatial aliasing, measurement noise, and normally distributed microphone positioning errors are simulated and assessed in different characteristics.

A plane wave sound field coming from direction $\Omega(\theta, \phi) = (90°, 2°)$ is simulated on a rigid sphere array with radius $r = 0.027\,\text{m}$ and maximum order $N_{max} = 5$, distributing 50 microphones on a Lebedev grid. The aliasing frequency is at $f_a = 10\,\text{kHz}$ for all configurations, except in the aliasing investigations where the radius is altered to achieve different characteristics of aliasing. All measurement errors are provoked as follows:

- Spatial aliasing is simulated by altering the radius of the array, consequently changing its upper frequency limit f_a (according to $N = kr$, see [218, p. 81]). Consequently, aliasing can be expected for frequencies above f_a.
- Measurement noise is simulated for different noise levels by convolving white noise on the simulated impulse responses, measured as SNR in dB. Note that the SNR relates to the noise floor and does not represent the actual SNR between signal and noise.
- Microphone positioning errors are simulated by adding an angle offset $\Delta\theta$, $\Delta\phi$ (in degrees) to the sample positions in azimuth and elevation direction, respectively. The offset is normally distributed over all samples.

Table 4.2 provides an overview of the listening test conditions.

Table 4.2: Listening test conditions for each error. Condition 1 is the reference (REF), condition 6 the anchor signal (ANC)

	Aliasing	Noise	Position Offset
Condition	f_a [kHz]	SNR [dB]	$\Delta\theta$, $\Delta\phi$ [^{o}]
1 (REF)	10	-	-
2	8	-90	0.009
3	6	-88	0.015
4	4	-84	0.02
5	2	-82	0.2
6 (ANC)	1	-20	0.5

All sound fields are simulated using the Sound Field Analysis toolbox (SOFiA) [30]. The ten test signals listed in Table 4.1 are presented to the assessors in a multi-stimulus listening experiment.

4.1.1 Listening Test Design

The listening test addresses each error category separately. It is divided into two parts: First, a training is conducted to familiarize the assessors with the audio signals and the extreme error conditions, in order to develop an understanding of the quality range as well as to improve the intra-rater reliability which describes the degree of stability observed when a measurement is repeated under identical conditions by the same assessor. After a short break, the actual listening experiment takes place which is based on the multi-stimulus with hidden reference and anchor (MUSHRA) paradigm. In MUSHRA, a certain number of test conditions is compared to an open reference. The conditions comprises SUTs, a hidden version of the reference (REF), and at least one anchor (ANC). However, the test design as originally proposed in [52] was altered and adapted to the aim of the present experiment: Instead of using a low-pass filtered signal as ANC, for each error category, condition 6 from Table 4.2 is used. Accordingly, assessors are explicitly instructed to rate ANC with a value corresponding to their trained inner reference.

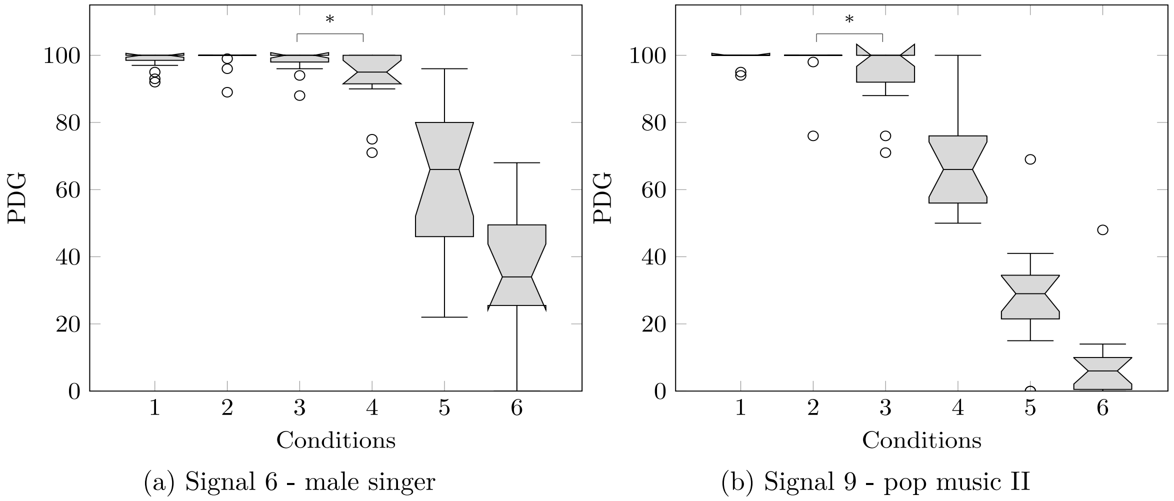
Assessors are instructed to rate all test conditions on a 100-point continuous scale in comparison to REF. In addition, REF is arbitrarily hidden within the test conditions and should be rated with PDG = 100. Moreover, the scale is equally divided into five quality categories according to the Mean Opinion Score (MOS) [199]: *excellent*, *good*, *fair*, *poor*, and *bad.* This metric should support assessors with additional anchors in case they are not sure how to distribute their quality impression onto the scale.

The listening test software enables real-time switching between test conditions without stopping audio playback. It is also possible to play the signals in a loop and to change the loop limits in order to zoom into a particular signal part for detailed analysis. In the experiment, all items and conditions are presented in random order. All in all, 30 items, comprising three errors for ten test signals, are assessed, whereas each item comprises six conditions: five degraded signals (including ANC) and the hidden REF.

4.1.2 Listening Test Results

Nineteen test persons with an average age of 26.4 years participated in the experiment, all stating to have normal hearing ability. They can be regarded as expert listeners as they are all familiar with the assessment of spatial audio systems in general, and of binaural auralizations in particular. None of the assessors seemed to have misrated systematically. Therefore, the ratings of all assessors are subject to the following analysis. A Kolmogorov-Smirnov-Test revealed normally distributed data. On average, the listening experiment took 38 minutes.

Spatial aliasing: Figure 4.2 exemplarily shows the PDGs for aliased sound fields for signals 6 and 9, the male singing voice and pop music, respectively. (Results for all other signals are given in Appendix D.)



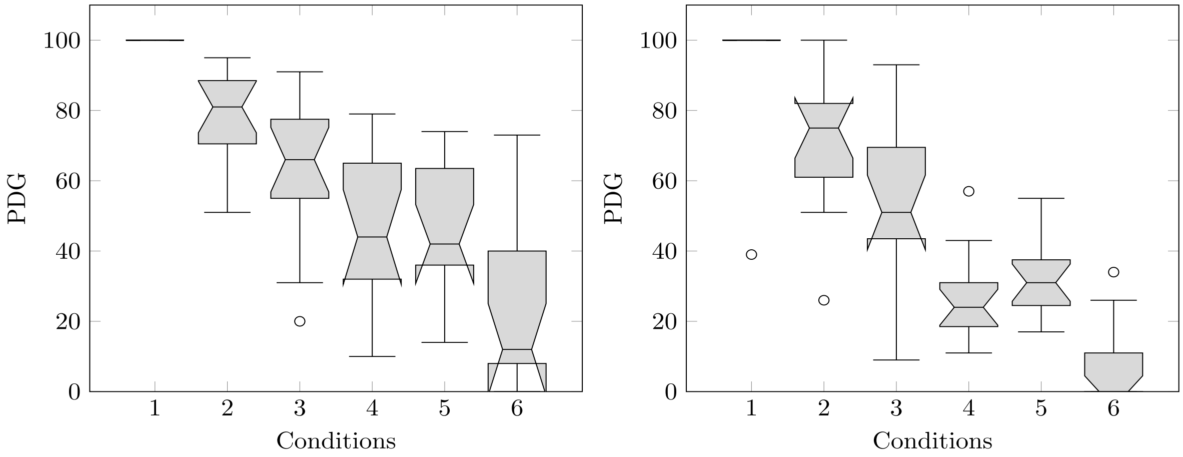
(a) Signal 6 - male singer

(b) Signal 9 - pop music II

Figure 4.2: Box plots showing PDGs for spatial aliasing exemplarily for signals 6 and 9. The median is marked with a black line and outliers with a circle.

Results indicate that spatial aliasing gradually decreases the auralization quality with increasing aliasing level. The effect is signal dependent which can be seen by comparing conditions 1-4 in both plots: For the male singer, aliasing seems to have only little or no effect at all, at least no significant, whereas the quality of the pop music signal is already significantly reduced for conditions 3 and 4. However, ratings are still in the excellent to good region. The signal dependent behavior is further underlined comparing condition 5 and 6. For signal 9, they are both significantly rated poor while aliasing for signal 6 ranges from fair to good, respectively. Also note the width of the confidence intervals for these conditions, which are represented by the size of the notches. They indicate a stronger agreement of the assessors when rating aliasing for the pop music signal than for the male singer. All other signals show a similar behavior with only slightly differing PDGs and confidence intervals. The respective plots are given in Figure D.1 in Appendix D. All in all, spatial aliasing has a stronger effect on the auralization quality for signals 2, 8, and 10, the clarinet, organ, and toms, where quality is significantly decreased from condition 3 on. For guitars, pop music, and orchestra, i.e., signals 4, 5, and 7, quality significantly degrades from condition 4 on. Also it can be summarized that an aliasing frequency limit at $f_a = 8$ kHz seems sufficient in practical applications, at least for the tested signals and an order $N = 5$ array. Moreover, even in the presence of small aliasing amounts, such as for conditions 3 and 4, PDGs are still in the excellent to good region.

Measurement noise: Measurement noise corrupting the sound field with noise levels according to Table 4.2 is addressed next. Figure 4.3 exemplarily shows the PDGs for signals 6 and 9, i.e., male singer and pop music:



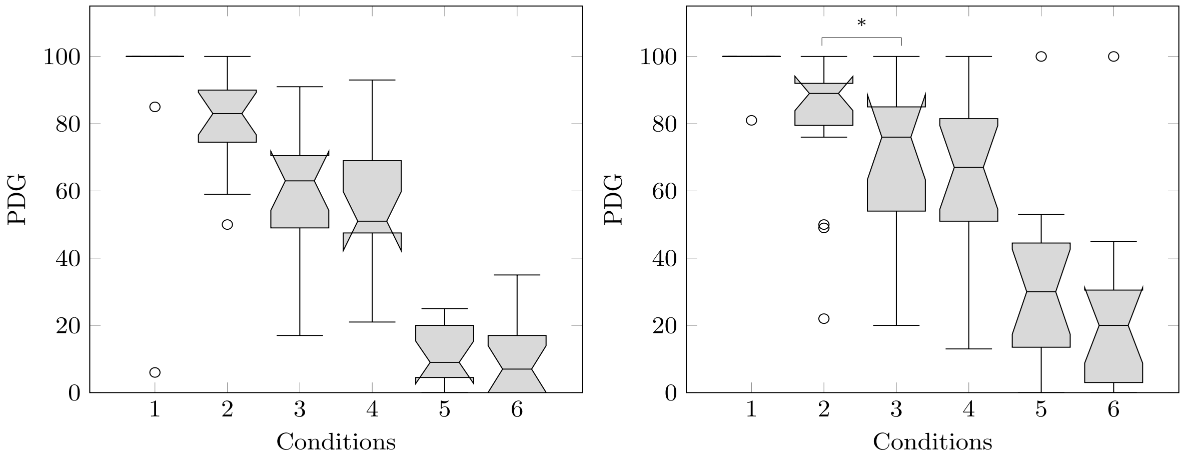
(a) Signal 6 - male singer

(b) Signal 9 - pop music II

Figure 4.3: Box plots showing PDGs for measurement noise exemplarily for signals 6 and 9. The median is marked with a black line and outliers with a circle.

As expected from theory (see Section 3.1.5), uncorrelated noise has a strong impact on auralization quality for all test signals. Starting with condition 2, with SNR = 90 dB, the reproduction quality is significantly reduced and decreases even further with increasing noise level. In case of aliased sound fields, the effect of noise is signal dependent and can clearly be seen by comparing conditions 4 and 5. Ratings for signal 6 are still in the fair region while for signal 9, ratings significantly drop to poor quality. In addition, quality is rated higher by trend for signal 9 while no difference was found between the same conditions for signal 6. Although this effect is not significant, it is still notable in the small confidence intervals, indicating an agreement of the assessors. This behavior can also be seen for signals 2, 4, and 5. Relating to condition 6, ratings significantly drop to bad quality, whereas this effect is stronger for signal 9 as indicated by the small box. Although results are slightly differing in PDG and confidence, all other signals show a similar tendency when uncorrelated noise is present. Respective plots for all other signals are given in Figure D.1 in Appendix D. It can be observed that a SNR of 90 dB leads to a significant quality reduction for all signals, with most ratings in the fair region or lower. Only signals 1 and 2, the castanets and the clarinets, are still rated good for condition 2. Higher noise levels, however, further decrease the reproduction quality and lead to unacceptable quality ratings, mostly in the poor to bad range.

Positioning errors: Figure 4.4 shows the PDGs for normally distributed microphone positioning errors. Plots a) and b) depict offset errors exemplarily for signals 3 and 6, corresponding to the female and male singer items, respectively.



(a) Signal 3 - female singer

(b) Signal 6 - male singer

Figure 4.4: Box plots showing PDGs for normally distributed microphone positioning errors exemplarily for signals 3 and 6. The median is marked with a black line and outliers with a circle.

Results show the gradual decrease in quality with increasing offset error. A normally distributed positioning error of 0.009°, represented by condition 2, leads to a significant quality reduction for both signals, with ratings still in the excellent to good region. Further increasing the offset to 0.006°, like in condition 3, further reduces the auralization quality significantly for both signals, whereas another 0.005° (condition 4) shows no significant impact at all. However, all ratings for conditions 2–4 are still in the good to fair range for both signals, until quality significantly drops into the poor to bad region for conditions 5 and 6 when further increasing the offset error to 0.2° and 0.5°, respectively. No significant differences can be seen between conditions 5 and 6 for both signals. REF is correctly identified in all cases. PDGs for all other signals are shown in Figure D.3 in the appendix. For these signals, results basically follow the same trend as the two examples presented above, with conditions 5 and 6 being rated worse in the poor to bad range. However, in most cases, where displacements are very small, ratings are significantly higher and are mostly between excellent to fair quality, which is strongly depending on the presented audio signal. For example, these two conditions significantly differ only for signal 9. For the remaining conditions 2–4, significant differences are observable for signals 7 and 9, while variations for all other signals are only present by trend.

4.1.3 Discussion

Experiment I assessed the influence of measurement errors on the reproduction quality for arrays simulated under free-field conditions using ten different test signals. Specifically, the experiment evaluated the impact of spatial aliasing, measurement noise, and normally distributed positioning errors for five gradually increasing error characteristics. Results showed that the impact of each error was slightly depending on the presented signal type, although its overall perceptual effect seemed similar for all signals. For spatial aliasing, the maximum aliasing frequency of all arrays was set to $f_a = 10\,\mathrm{kHz}$. Results indicate that a certain degree of aliasing is acceptable for all tested signals, at least for order $\mathrm{N}_{max} = 5$ arrays, since a significant reduction in quality appears only for aliasing frequencies below 8 kHz. For decreasing aliasing frequency, the corresponding quality was mostly rated from fair to bad. Measurement noise had a stronger impact on the auralization, significantly degrading the reproduction quality from condition 2 on for all tested signals. While for condition 2 (noise with $\mathrm{SNR} = 90\,\mathrm{dB}$), quality was still rated from good to fair, higher noise levels led to unacceptable quality ratings, all in the poor to bad range. The overall perceptual impact of noise was only slightly differing between test signals. Like spatial aliasing, microphone positioning offsets seem to be acceptable to a certain degree. For most signals, positioning errors of up to 0.02° led to a slight, yet significant decrease in quality, still being rated from excellent to good. Stronger offset errors resulted in significantly worse ratings in the poor to bad quality range, which are not acceptable in practical applications.

4.2 Experiment II: Quantitative Error Analysis in Reflective Environments

While Experiment I quantitatively evaluated system errors under free-field conditions, Experiment II investigates their impact in reflective environments, with the aim to reveal the influence of the reflection properties of the simulated environment on the perception of measurement errors. Particularly, a spherical microphone array is simulated in two shoe-box shaped rooms with similar dimensions, but with differing absorption characteristics, one with weak and the other with strong reverberation. Based on similar array configurations as used in Experiment I, spatial aliasing, measurement noise and normally distributed microphone positioning errors are assessed in a quantitative listening experiment. Also, the same ten test signals are used.

4.2.1 Room Simulations

All room simulations in Experiment II are realized using the MC-RoomSim MATLAB toolbox [287]. In the toolbox, the early sound field is modeled based on a mirror image source model (MISM) [143] while the late reverberation is modeled using temporally shaped white noise. The software also takes sound scattering and absorption at the boundaries into account. In this experiment, the reverberation behavior of the simulated environments is modeled by altering the absorption characteristics of the boundary elements, i.e., walls, ceiling, and floor. Specifically, the dry room is simulated with high and the reverberant room with low absorption coefficients, resulting in reverberation times of approximately $T_{60} = 0.4\,\mathrm{s}$ and $T_{60} = 4.5\,\mathrm{s}$, respectively. To keep the complexity of the experiment in manageable amounts, only a fixed set of frequency-dependent scattering coefficients are used for both simulated rooms. They are shown in Table C.1 in Appendix C. The sound field is then calculated for every microphone of the array, resulting in a set of RIRs which are subsequently auralized using SOFiA [30] for binaural synthesis. An omni-directional sound source and the array are positioned in the simulated rooms, both providing equal dimensions, with a volume of $1600\,\mathrm{m}^3$ and corresponding side proportions of $20\,\mathrm{m}$ length, $16\,\mathrm{m}$ width and $5\,\mathrm{m}$ height. Note that sound source and array are positioned in a non-symmetrical fashion to avoid directional biases in the synthesis, like, for example, front-back confusions. The Cartesian coordinates of the array are $[x_a, y_a, z_a] = [5, 8, 1.7]$ and $[x_s, y_s, z_s] = [17, 9, 1.7]$ (in meters) for the sound source, respectively.

Table 4.3: Listening test conditions for Experiment II listing array configurations, reflection properties of the room, and the tested error degrees for system errors.

		Aliasing	Noise	Position Offset
Room	Condition	f_a [kHz]	SNR [dB]	$\Delta\theta/\Delta\phi$ [^{o}]
Dry	1 (REF)	10	-	-
	2	8	-90	0.009
	3	4	-84	0.02
	4 (ANC)	1	-20	0.5
Reverberant	5 (REF)	10	-	-
	6	8	-90	0.009
	7	4	-84	0.02
	8 (ANC)	1	-20	0.5

4.2.2 Listening Test Design

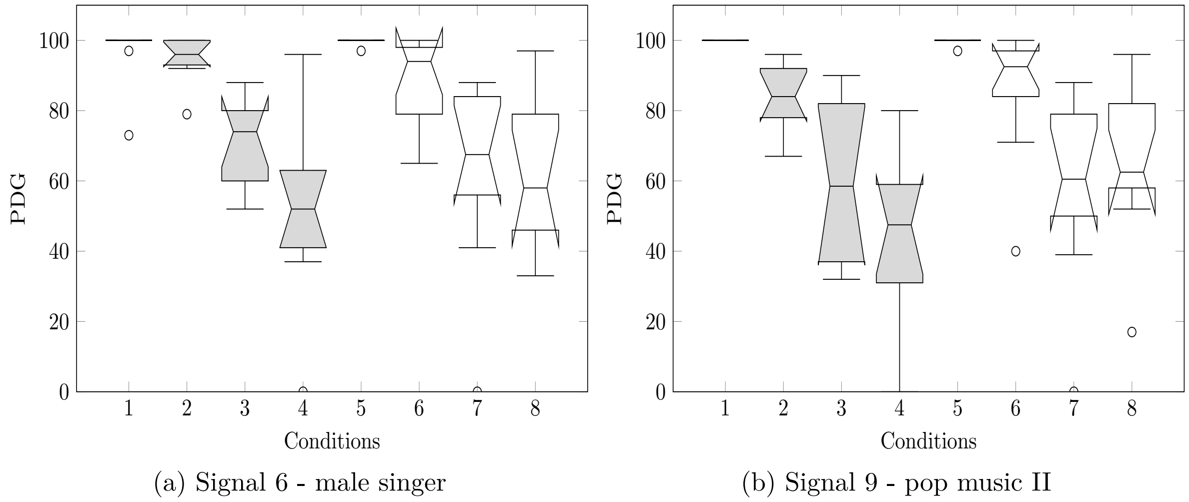
Similar to Experiment I, a modified MUSHRA test design is chosen for analysis. Again, all assessors are familiarized with the extreme error conditions in a short training session prior to the actual listening experiment. An order $N_{max} = 5$ spherical microphone array with radius $r = 0.027\,\text{m}$ is simulated in the two shoe-box shaped rooms, employing a Lebedev quadrature with 50 samples which are distributed on a rigid sphere. For each room, three different error characteristics are evaluated, corresponding to conditions 2, 4, and 6 in Table 4.2. Again, all ten test signals listed in Table 4.1 are presented in the experiment. Like in the free-field experiment, REF is hidden within the test conditions and ANC, as originally proposed in [50], is discarded and replaced by the strongest error characteristic. Table 4.3 gives an overview of all test conditions.

4.2.3 Listening Test Results

Note that the results of Experiment II, specifically of noise and offset errors, are somehow flawed. For this flaw, unfortunately, no satisfactory explanation could be found. Refer to the discussion in Section 4.2.4 for details.

The present analysis aims at revealing the influence of room reflections, identifying how they contribute to the perception of measurement errors. Eighteen test persons participated in the experiment with an average age of 25.7 years. All assessors are evaluated with a short questionnaire assessing their experience in listening experiments in general, in the evaluation of VAEs as well as logging known hearing issues. All participants stated to have no known hearing disabilities and can be regarded as expert listeners as they are all familiar in the assessment of spatial audio systems in general, and binaural auralizations of spherical microphone array data in particular. Including the training session, the listening experiment took 42 minutes on average to evaluate all 60 items, represented by three errors, ten test signals, and two rooms. For each item, assessors are asked to rate four conditions comprising the three error characteristics and the hidden REF. None of the assessors seemed to have misrated systematically. Therefore, the ratings of all assessors are subject to the following analysis. A Kolmogorov-Smirnov-Test revealed normally distributed data. Note that directly comparing system errors between both rooms is not possible based on the perceptual data derived, as both rooms were assessed in two distinct listening experiment sessions. Therefore, the room dependence can only be evaluated in relation to single conditions for a specific environment but not in terms of absolute PDG ratings by comparing both rooms.

Spatial aliasing: Plots a) and b) in Figure 4.5 show the PDGs for the aliased sound fields exemplarily for signals 6 and 9, the male singer and pop music II, respectively.



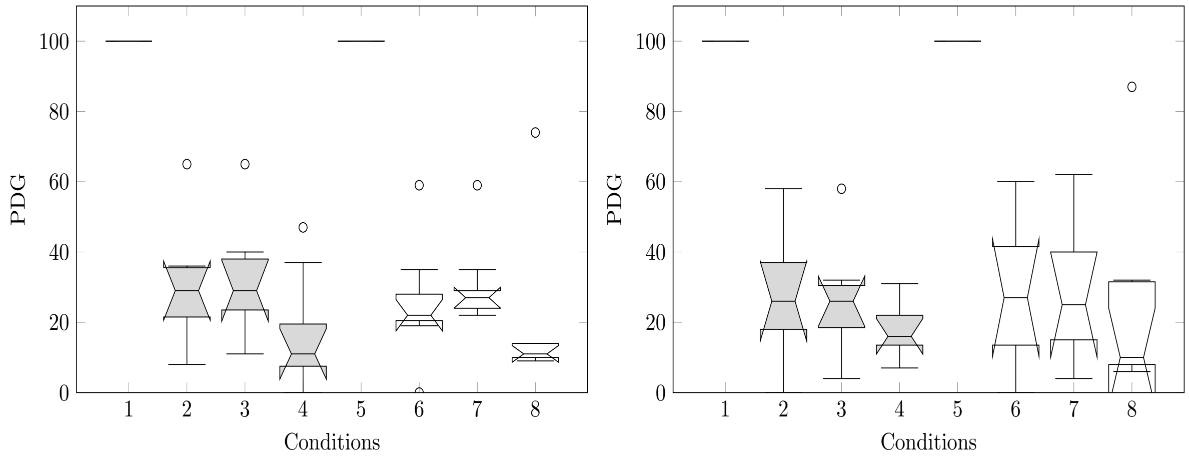
(a) Signal 6 - male singer

(b) Signal 9 - pop music II

Figure 4.5: PDGs for spatial aliasing exemplarily for signals 6 and 9 in simple room geometries. The gray boxes represent the dry, white boxes the reverberant room. Notches in the box plots indicate the 95 % confidence interval. The median is marked with a black line and outliers with a circle.

Conditions 1–4, the gray boxes, relate to the dry, and conditions 5–8 to the reverberant room (white boxes). The ratings indicate that aliasing gradually decreases the auralization quality with increasing error level in a similar fashion as in the free-field experiment described in Chapter 4.1. Comparing both plots, it can be seen that even small amounts of spatial aliasing lead to a significant reduction in quality, but with ratings still in the excellent region, like for conditions 2 and 6. Further reducing the aliasing frequency results in stronger artifacts which significantly corrupt the reproduction quality. However, ratings in the dry (conditions 3 and 4) and the reverberant room (conditions 7 and 8) are still in the good to fair range. Also, the overall trend seems similar for both signals, with only slight differences in the variance as indicated by the box sizes and the length of the whiskers. Results for all other signals are shown in Figure D.4 in Appendix D. Basically, they all follow the same trend as Signals 6 and 9, suggesting that the reverberation characteristics of the simulated environment have no impact (or only little) on the perception of spatial aliasing artifacts. Only slight differences can be seen between audio signals, like comparing signals 3 and 5 in the dry room. While for the first, each condition leads to a significant quality reduction, no significant difference can be observed at all for the latter. Although no direct comparison between both rooms is possible in terms of absolute PDG values, it can be seen for all signals that corresponding conditions are mostly rated similar and only slightly differ in their box sizes and whiskers. In addition, it is noteworthy that no ratings are given in the poor to bad quality area for all signals. This might be an effect of rating uncertainties due to their trained inner ANC impression, or possible methodical issues in the test design which could lead to wrong scale usage. In some cases, these uncertainties are also indicated by bigger boxes and extreme whiskers.

Measurement noise: Figure 4.6 exemplarily shows PDGs for signals 6 and 9, which are presented in plots a) and b), respectively. It becomes clear that the presented measurement noise has a stronger impact on quality than spatial aliasing, because all ratings are located in the poor to bad region. While the first two erroneous conditions seem to result in the same low quality, the strongest error conditions, conditions 4 and 8, are significantly degrading the quality, at least for signal 6. For signal 9, however, these quality differences are only observable by trend. REF is correctly identified in all cases. The described behavior can also be seen for all other signals which are plotted in Figure D.5 in Appendix D. Condition 4 is significantly rated worse than conditions 2 and 3 for signals 5, 7, and 8, while no significant differences are found for all remaining signals. For all signals, no significant differences are present between conditions in the reverberant room, all rated in the poor to bad quality range. Between rooms, no significant influence of the reverberation on noise perception is observed.

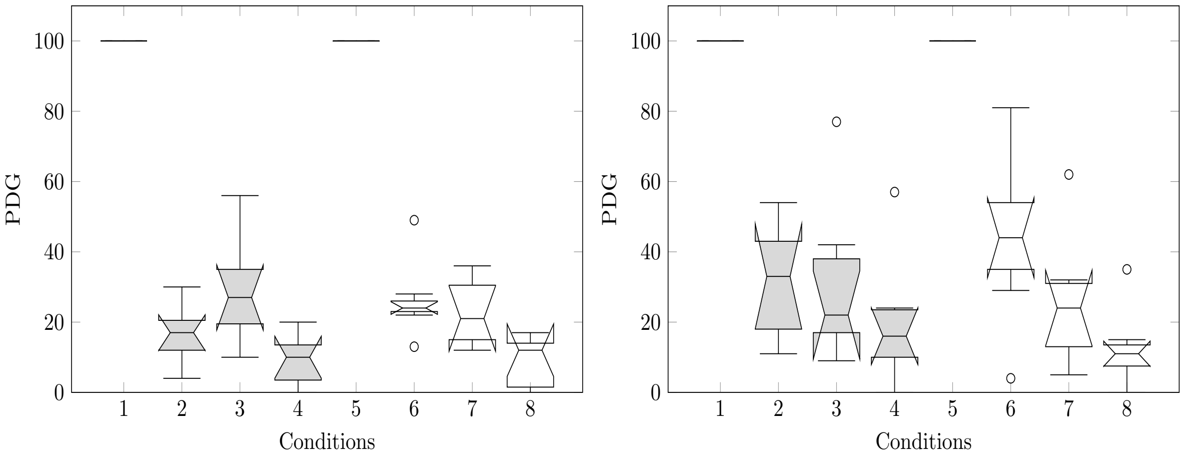


(a) Signal 6 - male singer

(b) Signal 9 - pop music II

Figure 4.6: PDGs for measurement noise exemplarily for signals 6 and 9 in simple room geometries. The gray boxes represent the dry, white boxes the reverberant room. Notches in the box plots indicate the 95 % confidence interval. The median is marked with a black line and outliers with a circle.

Positioning errors: Figure 4.7 exemplarily shows the PDGs for normally distributed microphone positioning errors (over azimuth and elevation direction) for signals 3 and 6 in plots a) and b), respectively. Overall, it can be seen that ratings for the female singing signal are in the poor to bad range in both rooms. In the dry environment, conditions 2 and 4 are rated worse while, interestingly, condition 3 is significantly rated higher. The behavior is different in the reverberant environment, with no significant difference between conditions 6 and 7, and condition 8 being significantly rated worse. Position errors for the male singing signal in plot b) show a gradual decrease in quality, which is tendentially in the dry, and significant in the reverberant environment. Also note that condition 6 is rated in the fair region while conditions 7 and 8 reduce the quality into poor and bad area. This indicates that the impact of positioning errors is dependent of the reflective properties of the simulated environment. PDGs for all other signals are shown in Figure D.6 in the Appendix D. The above mentioned effect that condition 3 significantly outscores conditions 2 and 4 can also be observed for signals 2, 5, and 7, while for signals 4, 8, and 10 it is only present by trend. No significant differences between the dry room conditions can be seen for signals 1 and 9, with ratings for the first still being in the fair to poor range. In the reverberant environment, an offset increment leads to a gradual decrease in quality for all signals, which is significant at least from condition 6 to 8. Results indicate the influence of reverberation on the perception of offset errors, which is also depending on the content of the test signal used.



(a) Signal 3 - female singer

(b) Signal 6 - male singer

Figure 4.7: PDGs for microphone positioning errors exemplarily for signals 6 and 9 in simple room geometries. The gray boxes represent the dry, white boxes the reverberant room. Notches in the box plots indicate the 95 % confidence interval. The median is marked with a black line and outliers with a circle.

4.2.4 Discussion

Experiment II evaluated the influence of measurement errors in simple room geometries for ten different test signals. In comparison to the results of Experiment I, system errors seemed to have a similar impact on the reproduction quality under reflective environments as in free-field sound fields. However, slight differences were observed which mainly depended on the presented audio signal. While spatial aliasing gradually decreased the reproduction quality with increasing error strength, noise and positioning errors seemed to have a stronger impact in reverberant environments than under free-field conditions, as quality for most signals was mainly rated from poor to bad. For positioning errors, noticeable differences between rooms were mainly observed for distinct conditions while only slight differences were present for aliased and noisy sound fields.

What strikes are the extreme scores for noise and positioning errors, specifically when comparing the same conditions to the free-field simulations in Experiment I; possible causes are discussed next. First, the extreme scores are actually the result of noise and positioning errors in reflective environments, indicating that even a little amount of noise (or microphone position offset) which is acceptable in free-field environments significantly deteriorates the reproduction quality. Second, the extreme scores point out methodical problems in the listening experiment design. Recall that all conditions in this experiment were a subset of conditions from Experiment I, but with two intermediate error levels missing. The lack of these quality conditions might have led to rating uncertainties as well as biased scale usage by the assessors. In contrast, however, such rating uncertainties were not observed for spatial aliasing in reflective envrionments. Third, the extreme scores might also point to flaws in the simulation parameters. This, however, seems unlikely as the same simulation and auralization tools (and parameters) were also applied for spatial aliasing as well as the room simulations conducted in Experiment III. Finally, It should be noted that all assessors stressed having problems to rate the quality of a stimulus when comparing poor sounding signals against each other. This rating uncertainty is likely to have biased the results. Nevertheless, even if the extreme scores of noise and positioning errors are subject to flaws in the simulations or the listening test, the main goal of this thesis, enbling quality prediction, is still achievable as will be shown in Chapter 5. In this sense, the extreme scores for noise and positioning errors in reflective environments are assumed to be correct.

In sum, however, it can be stated that the environmental reflection properties seem to have no, or only little, impact on the perception of measurement errors in spherical microphone arrays.

4.3 Experiment III: Descriptive Quality Analysis based on ASW and LEV

While Experiments I and II assessed the influence of measurement errors in quantitative analyses, Experiment III focuses on the qualitative assessment of the reproduction quality using ASW and LEV as quality factors as they are common in the assessment of concert halls. In this experiment, the synthesis quality is assessed for various array configurations employing different sound field orders. The data fundamental to the following analysis is based on spherical array simulations in three environments with varying reflection characteristics. Specifically, the same environments are simulated as used in Experiment II: a free-field sound field and two shoe-box shaped rooms, one with weak, the other with strong reverberation.

4.3.1 Listening Test Design

Three different spherical arrays employing maximum orders $N_{max} = 3$, 5, and 8 are simulated in three different environments: free-field, a dry room, and a reverberant room. The two rooms had the same dimensions and absorption properties as described in Section 4.2. Like in Experiments I and II, the aliasing limit of all arrays is set to $f_a = 10\,\text{kHz}$. The listening test is based on a multi-stimulus design using the free-field simulations as ANC, because it is expected to provide the lowest spatial quality. The $N_{max} = 8$ array in the reverberant room serves as REF. Table 4.4 lists all conditions used in Experiment III.

Table 4.4: Listening test conditions fundamental to Experiment III, comprising the array configurations used in terms of radius, used sound field order, number of samples, and the reflection properties of the simulated environments.

		Radius	Sound field order	Samples
Room	Condition	r [m]	N_{max}	Q
Free-field	1 (ANC)	0.016	3	26
Dry	2	0.016	3	26
	3	0.027	5	50
	4	0.044	8	110
Reverberant	5	0.027	3	26
	6	0.027	5	50
	7 (REF)	0.044	8	110

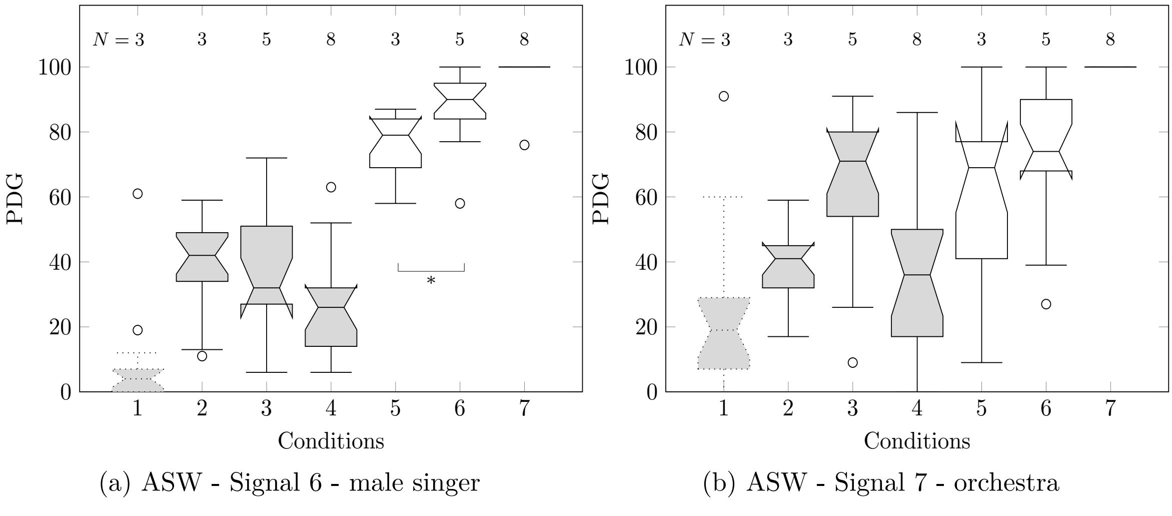
Again, the experiment is divided into a training session in which all assessors are familiarized with extreme examples of ASW and LEV, and the actual listening experiment. Assessors are prompted to rate ASW and LEV for all conditions on a 100-point scale in comparison to REF which is also hidden within the test conditions. Again, the ten test signals listed in Table 4.1 are used as test material.

4.3.2 Listening Test Results

Twenty-two assessors with an average age of 28.3 years participated in the listening test. All can be regarded as expert listeners as they are all familiar with listening experiments and, in particular, with the evaluation of binaural auralizations. In addition, 13 assessors specifically had experience in the assessment of ASW and LEV. Furthermore, all stated to have no known hearing disabilities. None of the assessors seemed to have misrated systematically. Therefore, the ratings of all assessors are subject to the following analysis. A Kolmogorov-Smirnov-Test revealed normally distributed data.

ASW

Figure 4.8 exemplarily shows ASW ratings for signals 6 and 7, the male singing voice and the orchestra recording, respectively.



(a) ASW - Signal 6 - male singer

(b) ASW - Signal 7 - orchestra

Figure 4.8: ASW ratings as notched box plots for signals 6 and 7. The gray box with the dotted line indicates ANC, while gray and white boxes with solid lines represent the dry and the reverberant room, respectively. The median is marked with a black line, outliers are marked with a circle.

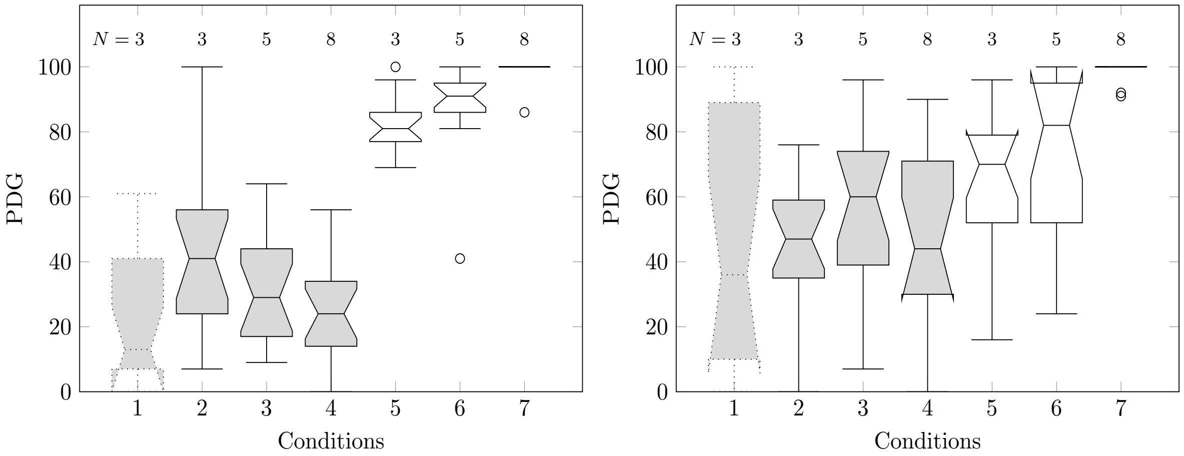
Note that, in the following analysis, the box-plots representing subject ratings for the three reflective environments (the free-field sound field as well as the dry and the reverberant room) are coded using different gray-scales and box-enclosing line-styles: ANC is represented by the gray box with the dotted line, the dry room conditions by the gray box with the solid line, and the reverberant room by the white boxes.

The signal-dependency when assessing ASW becomes obvious by comparing plots a) and b). For condition 1, the free-field array, PDGs show that the male singing voice was clearly identified as ANC, which is indicated by the small confidence intervals. Ratings for the orchestral recording on the other hand show larger confidence intervals, and whiskers ranging higher than PDG = 50, indicating that condition 1 could not confidently be identified as ANC. However, ASW ratings rise significantly when room reflections are present (conditions 2–7). Under these conditions, the array order also affects ASW perception. In the dry room, for signal 6, an increase from order 3 to 5 (condition 2 to 3) shows no significant effect while ASW rises significantly for signal 7. An interesting behavior can be observed when further increasing the order to $N_{max} = 8$, i.e., from condition 3 to 4: ASW ratings drop for both signals. This effect is significant for signal 6, comparing condition 2 and 4, and for signal 7 from condition 3 to 4. This notable phenomenon is discussed in more detail in Section 4.3.3. In contrast, in the reverberant room, represented by conditions 5–7, ASW significantly rises with increasing sound field order. However, for signal 7, ASW is rated similar for condition 3 and 5, meaning that an order 5 array in a dry room provides the same ASW as an order 3 array in the reverberant room, at least for the orchestral signal. The reference condition REF was correctly identified for both signals.

ASW ratings for all other signals are presented in Figure D.7 in the appendix. Here, the general tendency as observed for signals 6 and 7 is similar for all other signals: ASW is rated lowest in the free-field situation and is rising with increasing reverberation. In the dry room, it is rated higher when the array order is increased from 3 to 5. This effect, however, is signal dependent and significant for signals 1, 3, 4, 5, and 9, and tendentially for all other signals. Increasing the order further to $N_{max} = 8$, surprisingly leads to a reduction of ASW for most signals by trend. For this behavior, significance can be observed for signals 1, 4, 5, and 9. However, highest ASW ratings are achieved in the reverberant room, represented by conditions 5–7. In these cases, the perception of ASW increases for all signals with increasing sound field order, except for signal 4, the guitars, where orders 3 and 5 show no difference in ASW perception, and only order $N = 8$ leads to a significant increase of ASW perception.

LEV

In order to illustrate the influence of the sound field order, the reverberation, and the test signal type on LEV perception, Figure 4.9 exemplarily shows PDGs for signals 6 and 7.



(a) LEV - Signal 6 - male singer

(b) LEV - Signal 7 - orchestra

Figure 4.9: LEV ratings as notched box plots for signals 6 and 7. The gray box with the dotted line indicates ANC, while gray and white boxes with solid lines represent the dry and the reverberant room, respectively. The median is marked with a black line, outliers are marked with a circle.

A clear signal dependency can be observed comparing plots a) and b). While PDGs for signal 6 are quite confident with small whiskers and boxes, clearly indicating the influence of the room and sound field order, ratings for signal 7 are more unclear, which is indicated by the whiskers nearly spanning over the whole scale, which can specifically be seen for condition 1 for signal 7 in plot b) of Figure 4.9. REF is clearly identified for both signals. In the dry room, represented by conditions 2–4, an interesting behavior can be seen towards less LEV when increasing sound field order. This effect is significant between conditions 2 and 4. For signal 7, no significant change in LEV perception is observed, which stands in contrast to the perception in the reverberant room. Here, LEV increases for both signals with increasing sound field order, which is significant for signal 6 and tendentially for signal 7. All other LEV ratings are shown in Figure D.8 in the appendix. Like in signal 7, it can be seen that unclear LEV ratings, represented by large whiskers and boxes, are also present for signals 1 and 2, the castanets and the clarinet signal, although not as extreme as for signal 7. Ratings for all other signals are more clear. PDGs indicate that the amount of reverberation affects LEV perception in a way that stronger reflections lead to higher LEV ratings, as is expected from concert hall acoustics theory [54]. This is significant for signals 3, 4, 5, 8, 9 and 10, and by trend for signals 1 and 2. The influence of the sound field order on LEV perception seems signal- and reflection-dependent. In the dry room, no significant differences are noticeable between LEV ratings for signals 1, 4, 5, 9, and 10, and in the reverberant room for signals 4, 5, and 10, at least for conditions 5 and 6. REF is significantly rated highest for all signals.

Three effects are worth noting: First, just like ASW ratings in the dry room, an order increment from 5 to 8 leads to significantly reduced LEV for signals 3 and 8, and by trend for signal 9. Second, several LEV ratings, like for signals 3, 6, and 8, show a similar trend as for ASW. And third, in LEV assessment, the identification of ANC seemed more difficult as in ASW evaluations, which can clearly be seen for signals 1, 2, 3, 5, and 8 where ANC had the same LEV ratings as orders 3 and 8 in the dry room. These effects are discussed in more detail in the next section.

4.3.3 Discussion

In the following, results of Experiment III are discussed. In general, it was observed that stronger reverberation led to higher ASW and LEV impression, which is in accordance to the findings from concert hall acoustics. Also, as would be expected from spherical arrays literature [8, 28], the sound field order significantly influenced the perception of ASW and LEV for most signals, and at least by trend for the other signals.

An interesting effect was observed in the dry room when increasing the sound field order from 5 to 8, resulting in a decrease of ASW and LEV perception. On the one hand, this would be expected from theory since higher orders lead to an increased spatial resolution of the sound field representation and therefore to a sharper source impression. On the other hand, this effect was not observed when increasing the order from 3 to 5 which resulted in higher ASW and LEV ratings. This effect occurred, at least by trend, for all signals in the dry room, but for none in the reverberant room, suggesting that the perception of ASW and LEV depends on an interaction of the sound field order employed and the reflection properties of the simulated environment. Based on the results so far, it is assumed that a higher spatial accuracy of the array leads to a more accurate representation of the room's reflection characteristics, consequently resulting in an improved and more realistic spatial impression. In addition, the strength of this effect was found to be signal dependent. (Note that this effect was also accurately estimated in another experiment described in [194] when applying a model for ASW and LEV estimation [137].) These results indicate that the observed behavior is not just an artifact of the test design, but actually a result of the characteristics in the data.

Although it has been shown in the literature that ASW and LEV are related in a way that an increased ASW leads to increased LEV perception and vice versa, like in [38, 195], the similar trend when comparing the ratings for ASW and LEV raises the question whether both percepts were correctly and distinctly assessed by the assessors or not. This specific question was accounted for in an interview with the assessors between training and actual listening test, in which all assessors stated that they were able to confidently separate both senses. Also, in the authors opinion, both impressions can clearly be perceived and assessed separately, especially after appropriate training. However, it seems that LEV was harder to assess than ASW as the size of the notches in the box plots and the larger whiskers indicate. REF and ANC were identified correctly for all signals, although the rating of ANC was not explicitly asked in the listening experiment. However, the least spacious items (free-field conditions and order 3 sound fields) always achieved the lowest ASW and LEV ratings.

5 Predictive Modeling

The previous chapter presented a perceptual analysis of spherical microphone array auralizations, investigating the influence of different array configurations and measurement errors in environments with varying reflection properties based on quantitative and qualitative analyses. In the following experiments, the resulting perceptual data and the output of the auditory model are used for predictive modeling. The objective is to identify and categorize system errors and predict their strength, estimate the reproduction quality in terms of ASW and LEV, and identify characteristics of the recorded sound field such as information about the room's reflection properties or the array configuration used.[11] This is achieved by applying models for classification and regression tasks. The evaluations are structured into three modeling experiments:

1. System errors are addressed based on data retrieved in Experiments I and II. A classification analysis is conducted to categorize the data into respective error classes. A subsequent regression approach estimates the strength of an error represented by ratings from the listening tests. After that, models trained with data from Experiment I are applied to predict the error ratings of Experiment II in order to see how such models would perform in practical applications. This way it is possible to estimate the influence of the reflection properties of the VAE on error perception which could not be evaluated in the listening test.

2. The quality of the reproduction is evaluated based on data derived in Experiment III. ASW and LEV are subject to a regression approach to predict their strength by estimating the ratings from the listening test.

3. An additional classification task assesses the reverberation of the recorded environment and the array configuration used, based on the data from Experiment III.

For all modeling experiments, the predictor importances are evaluated and discussed with the aim to identify MOVs predominantly contributing to the prediction performance.

[11] Parts of this analysis have been published in [194], or have been submitted for publication [192].

In predictive modeling, the so-called *no free lunch theorem* [302] states that no single model will perform better than any other model without substantial information of the problem to be modeled [301]. It is recommended to test and compare multiple models and then focus on the most promising in terms of prediction accuracy. Therefore, a number of models are applied for prediction and compared regarding their achieved accuracy and error scores. In the following experiments, the selected prediction techniques include simple models, like linear and partial least squares as well as more complex nonlinear models such as ANN, KNN, and SVM, and also decision trees are applied for prediction. All models are designed, trained, tuned, and tested using the statistics toolbox *R* [12]. Table 5.1 gives an overview on these models and the required packages.

Table 5.1: Overview of models used in the experiments for regression and classification, showing also the *R* packages[a].

Abbrev.	Model	Application	R package
LM	Linear model	Regression	*stats*
LDA	Linear discriminant analysis	Classification	*sparseLDA*
RLM	Robust linear model	Regression	*MASS*
RDA	Regularized discriminant analysis	Classification	*rda*
GLM	Generalized linear model	Regression	*MASS*
PLS	Partial least squares	Regression/Classification	*pls*
ANN	Artificial neural network	Regression/Classification	*nnet*
MARS	Multivariate adaptive regression splines	Regression	*earth*
FDA	Flexible discriminant analysis	Classification	*earth*
SVM$_{lin}$	Support vector machine (linear kernel)	Regression/Classification	*e1071*
SVM$_{rad}$	Support vector machine (radial kernel)	Regression/Classification	*e1071*
KNN	K-nearest neighbor	Regression/Classification	*RWeka*
BT	Bagged tree	Regression/Classification	*ipred*
RF	Random Forest	Regression/Classification	*random forest*
SGB	Stochastic gradient boost	Regression/Classification	*gbm*

[a] The core packages used for predictive modeling are *AppliedPredictiveModeling* and *caret* required. See [141, 173] for further details on predictive modeling in *R*.

As discussed in Section 3.3.3, a number of methods for data preprocessing can be applied to optimize the model performance. In the following analyses, multiple preprocessing and feature selection steps are conducted: The skewness of the data is evaluated and, if present, corrected with a Box-Cox transformation. Then, all predictors are tested against their variance and identified near-zero-variance predictors are subsequently removed from the data. Also linear combinations within the predictors are evaluated to identify collinearity. If linearly combined predictors are found, then they are also discarded

[12] The R Project for Statistical Computing, URL: www.r-project.org/

from the modeling process. Additionally, the correlation of predictors is measured and the results used to remove highly correlated predictors. In some cases, a PCA is applied for feature reduction, and the resulting components are further used for modeling. However, as noted in [141], these preprocessing steps do not necessarily optimize the model performance and sometimes even increase the risk of over-fitting, or the likelihood of selection bias—especially for small sample sizes. Therefore, the prediction accuracy is always tested and compared with different predictor sets, also including the full set. At the end of each analysis, a predictor importance assessment is conducted to provide information on what predictors contributed how strong to the model performance.

5.1 Error Prediction

This section deals with the prediction of system errors based on models for classification and regression. While the classification approach identifies and categorizes system errors, the subsequent regression analysis allows the quantification of the error strength by estimating the ratings from the listening experiments. In the first part of this section, both analysis approaches are conducted separately for Experiments I and II. The second part presents a further experiment to evaluate the model's prediction performance. Specifically, models trained with the free-field data from Experiment I are applied to predict the results from Experiment II, aiming to evaluate how these models perform on unknown samples, and to establish a relation between Experiments I and II.

Predictor prescreening

Before modeling, all MOVs are analyzed with respect to their inter-relationships, evaluating their variances, collinearities, and correlations. For this purpose, the predictors from Experiment I and II are combined and evaluated together in a single prescreening. First, a variance analysis reveals no near-zero-variance predictors. Second, collinearity is evaluated, showing that all spatial MOVs (ITDs, ILDs, and IACCs) from the basic and advanced version of PEAQ are linearly related. Consequently, the corresponding MOVs from the basic version are discarded from further analysis. The third step evaluates between-predictor correlations, identifying several highly correlated predictors as shown in the correlation matrix in Figure 5.1. Correlations are marked on a gray scale with black indicating high correlation while low correlations are brighter, with white representing zero correlation. Strong correlated pairs of MOVs with values above 0.90 are for example V1 and V2, V15 and V7, V7 and V25, V16 and V18, V17 and V19, V20 and V22, and

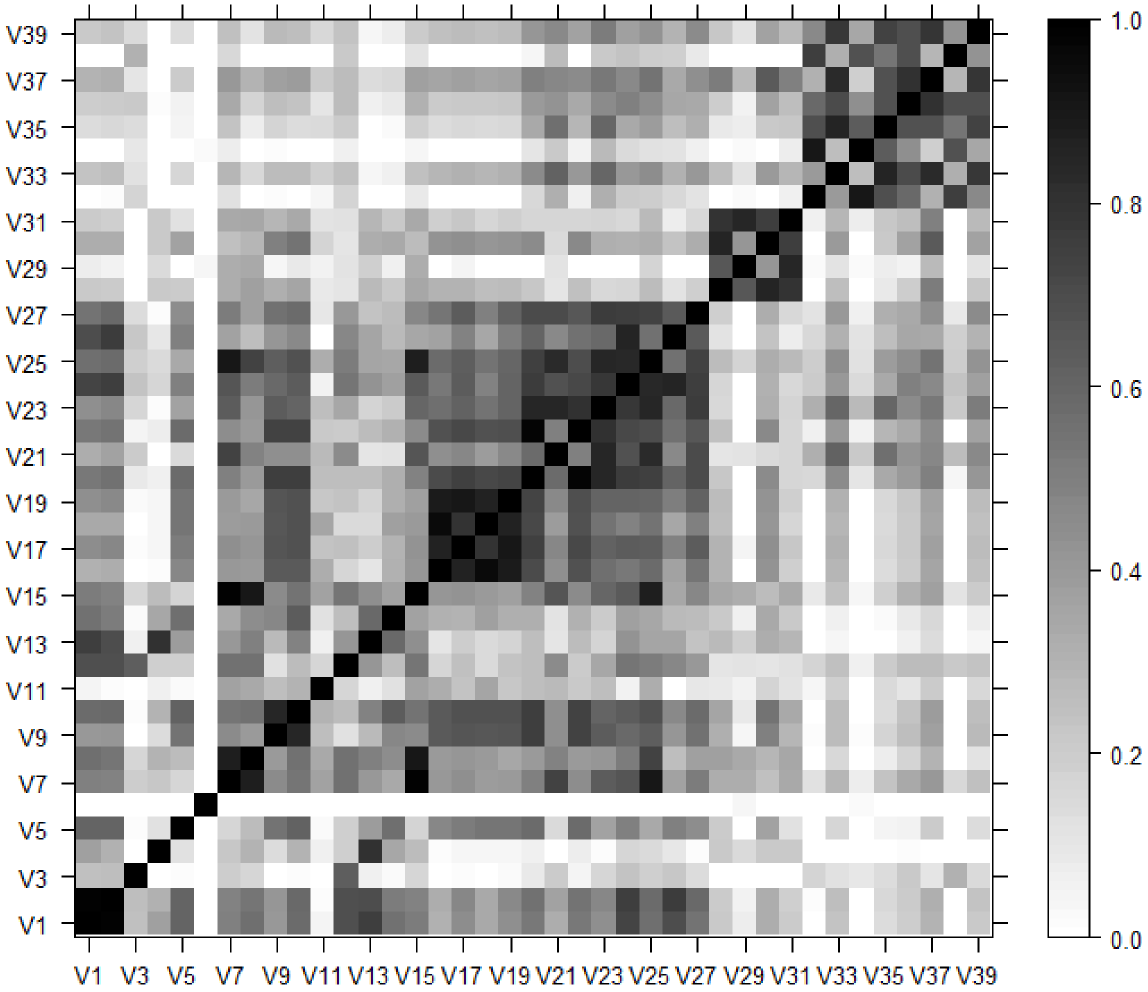


Figure 5.1: MOV correlation matrix showing between-predictor-correlations.

V32 and V34 (see Table 3.2 for an overview of MOVs). If need be, these predictors are removed and the remaining used as a subset for modeling, especially for models which are sensitive to highly correlated variables, like LMs. The prediction performance is then compared to the accuracy of the full feature set. The fourth step evaluates the skewness of the data which, if present, will be removed. Figure 5.2 shows the skewness statistics over all MOVs as a bar plot. Recall that values near zero indicate a symmetric distribution, positive values right, and negative values left skewness. Strong right-skewed distributions are observed for several MOVs, like V3, V34, or V38, and subsequently removed based on a Box-Cox-Transformation. In a last analysis step, a PCA is applied on the data to reduce dimensionality and identify predictors important for system error description.

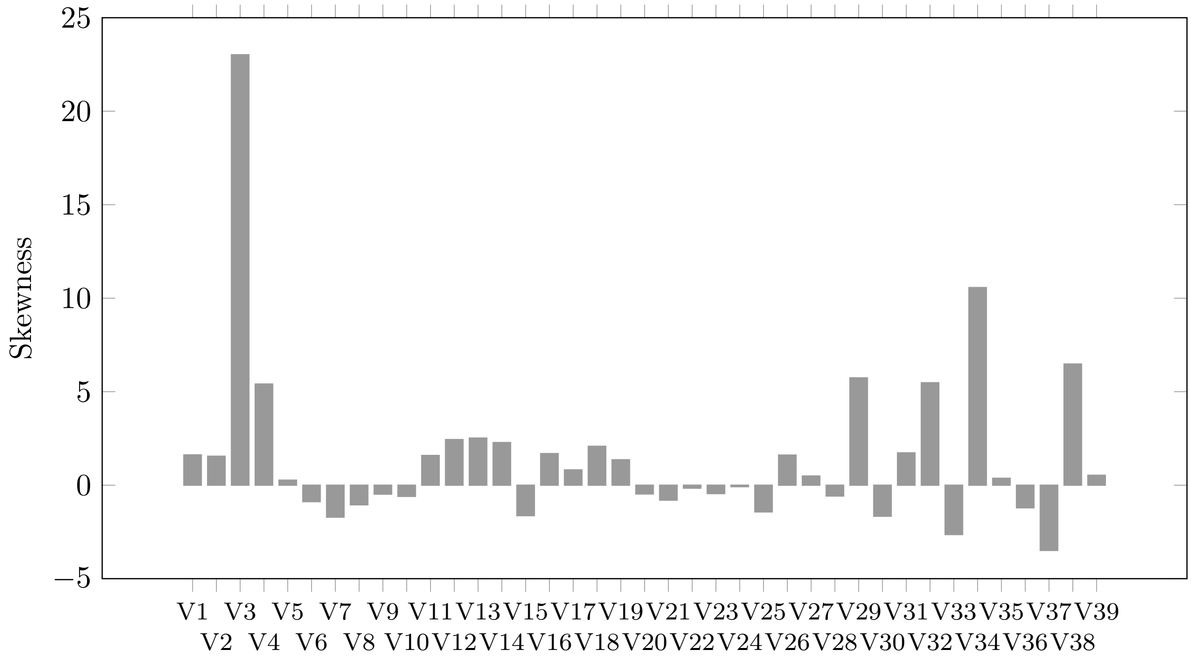


Figure 5.2: Predictor skewness over all MOVs for error classification in free-field.

All in all, five PCAs are conducted in this error prediction experiment: The first is based on all error data for all 70 test signals from SQAM while the remaining four PCAs evaluate each error separately. The data is centered and scaled prior to each transformation. Figure 5.3 exemplarily shows the explained variances of the first 11 components as bar plots, describing system errors in free-field and reflective environments. The first Principal Component (PC) comprises nearly 61 % of the information, the second about 26 %, and the third nearly 8 %, together explaining more than 95 % of the underlying information. So in practice, these three components should suffice to discriminate between the three error types: spatial aliasing, transducer noise, and randomly distributed microphone positioning offsets. In order to evaluate how strongly individual MOVs contribute to the corresponding components, their loadings can be viewed in a component matrix which is shown in Table C.4 in Appendix C. Only loadings above 0.30, which is an arbitrarily chosen threshold, are taken into account. These are the two bandwidth MOVs *BandwidthRef* and *BandwidthSut* as well as the RMS-modulation differences *rmsModDiff*, loading on the first component with 0.78, 0.50, and 0.32, respectively. PC2 can mainly be described by the average modulation differences *avgModDiff2*, the SUT-bandwidth *BandwidthSut*, and the RMS-modulation differences *rmsModDiff* with corresponding loadings of 0.31, 0.49, and 0.77. On the third component PC3, mainly the bandwidth measures *BandwidthRef* and *BandwidthSut* load with 0.59

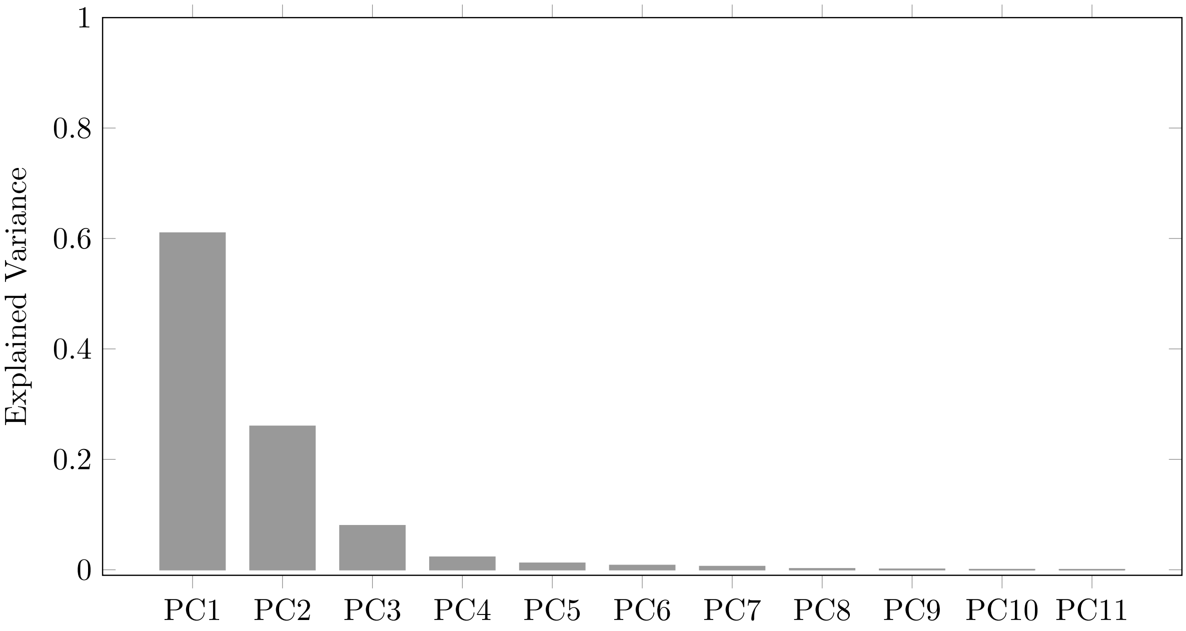


Figure 5.3: Explained variances for the first 11 PCA components for error classification.

and 0.70, respectively. The presented results of the PCA account for all errors of interest, for example in the classification experiment. However, to get a more detailed view on what predictors contribute to the description of individual errors, the following PCAs distinctly evaluate each error (The Tables C.5 to C.7 in Appendix C show their loadings). For spatial aliasing, only two components are needed to cover 95 % of the variance, with the bandwidth measures *BandwidthRef* and *BandwidthSut* strongly contributing with 0.68 and 0.67 on PC1 while the average modulation difference and the RMS-modulation difference *avgModDiff2* and *rmsModDiff* load with 0.37 and 0.85 on PC2, respectively. For measurement noise, three components account for 95 % of the information. In addition to the bandwidth measures *BandwidthRef* and *BandwidthSut*, providing loadings of 0.78 and 0.41, also the RMS-modulation difference *rmsModDiff* contributes with 0.41 to PC1. On PC2, The average modulation difference *avgModDiff2*, the bandwidht of the SUT *BandwidthSut*, and the RMS-modulation difference *rmsModDiff* load with coefficients of 0.34, 0.53, and 0.72, respectively, while the average modulation difference *avgModDiff2* as well as both bandwidth measures *BandwidthRef* and *BandwidthSut* load on PC3 with 0.42, 0.54, and 0.70, respectively. Also randomly distributed offset errors can confidently be described using three components: The MOV contribution is similar to noise but with slightly differing loadings. It can be observed that the bandwith measures *BandwidthRef* and *BandwidthSut* as well as the RMS-modulation difference *rmsModDiff* load on PC1 with 0.79, 0.50, and 0.30, respectively. On PC2, the average modulation difference *avgModDiff2*, the bandwidth measure for SUT *BandwidthSut*, and the RMS-modulation difference *rmsModDiff* contribute with loadings of 0.30, 0.46, and 0.79, respectively. Finally, PC3 is represented by the average modulation difference *avgModDiff2* as well as both bandwidth measures *BandwidthRef* and *BandwidthSut*, respectively contributing with loadings of 0.33, 0.56, and 0.70.

Note that the first three components mostly rely on the same MOVs for all errors, only slightly differing in their loadings: the modulations differences *rmsModDiff* and *avgModDiff2* as well as both bandwidth measures *BandwidthRef* and *BandwidthSut*. Although all these predictors seem important for error description, in practice a combination of MOVs might yield a better fit. This relationship, however, is not accounted for with a PCA and can only be evaluated by analyzing model-dependent predictor importances.

5.1.1 Classification in Free-Field Environments

In this classification task, multiple predictive models are applied to identify and classify system errors und free-field conditions. The classification process is different from regression insofar as no perceptual data is needed for prediction. Instead, four classes are predefined for this particular case, serving as observations in the modeling task. Class 1 is the error-free sound field, class 2 spatial aliasing, class 3 measurement noise, and class 4 represents normally distributed microphone positioning offsets.

All in all, 1260 samples representing the observations are available for analysis. All 70 test signals from the SQAM data base are convolved with the stimuli describing all three errors of Experiment I, each with 6 conditions (Table 4.2). All 39 MOVs serve as predictor variables in the following evaluations.

Model building and tuning

Before prediction, each model is trained with the training data set, comprising 80 % of the data. The remaining samples are held out for model testing. A generalization of the model performance is possible based on resampling, indicating how the models would perform when predicting unknown data. In this experiment, the 10-fold cross-validation is chosen for resampling, and repeated 50 times to increase the variance.

In order to find the best configuration for each model, parameter tuning is applied. This process choses the optimal tuning parameters based on the provided accuracy estimates in terms of the overall accuracy OA. Alternatively, also Kappa and RMSE

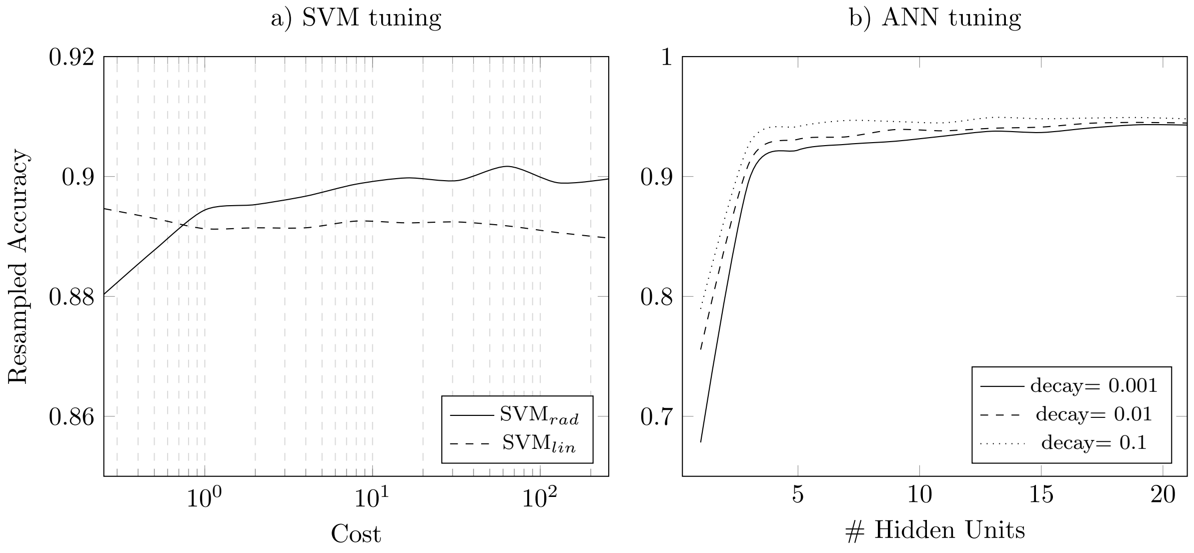


Figure 5.4: Estimated accuracies over tuning parameters for SVMs in plot a) and ANNs in plot b), for error classification based on 10-fold cross validation.

values are available to tune the models. Model tuning is exemplarily presented for a SVM with different kernel functions, a radial and a linear kernel, SVM$_{rad}$ and SVM$_{lin}$, and ANN using different weight decays for an increasing number of hidden units. Both SVMs are tuned using their cost parameter. Figure 5.4 shows the changing prediction accuracy over their respective tuning parameters for SVMs in plot a), and for the neural network in plot b). For example, the highest accuracy for SVM$_{rad}$, OA = 0.91, can be found at a cost value of 64 while SVM$_{lin}$ performs best at cost = 0.25. In case of the ANNs, best results are achieved using 21 hidden units and a decay of 0.1. This tuning process is conducted for all models to find the optimal tuning parameter settings. The best performing models are subsequently chosen for further evaluation.

Table 5.2 provides an overview of the models used in this classification task, their estimated accuracies, the chosen tuning parameters, and all applied preprocessing steps. All optimally tuned models show high classification accuracy estimates, mostly in the range of OA = 0.88–0.95. Best results are achieved with ANN, scoring 0.95, followed by SGB, RF, and FDA with values of 0.93, 0.93, and 0.92, respectively. Also LDA, SVM$_{rad}$ and KNN perform well, all providing accuracies of 0.90. PLS, RDA, and SVM$_{lin}$ show performances slightly below OA = 0.90, respectively with 0.89, 0.88, and 0.89. The lowest accuracy is estimated for BT, still reaching an accuracy of 0.79. However, these estimates seem—at least partly—quite optimistic. The model performances when predicting new samples are evaluated next, based on the hold-out test data set.

Table 5.2: Error classification performance, model parameters, and preprocessing steps.

Model	OA	*Kappa*	Optimal tuning parameter	Data preprocessing
LDA	0.90	0.88	-	BoxCox, C&S
PLS	0.89	0.86	ncomp = 19	BoxCox, C&S
RDA	0.88	0.85	-	BoxCox, C&S, PCA
ANN	0.95	0.93	size = 21, decay = 0.1	BoxCox, C&S
FDA	0.92	0.90	degree = 4, nprune = 30	BoxCox, C&S
SVM_{lin}	0.89	0.86	cost = 0.25	BoxCox, C&S, PCA
SVM_{rad}	0.90	0.87	cost = 64, sigma = 0.06357128	BoxCox, C&S, PCA
KNN	0.90	0.87	$K = 5$	BoxCox, C&S, PCA
BT	0.79	0.77	mincriterion = 0.5	
RF	0.93	0.91	mtry = 20	
SGB	0.93	0.91	ntrees = 150, depth = 3, shrinkage = 0.1, minobsinnode = 10	

Evaluation based on test data

In order to see how these models perform on unknown data, all models are evaluated against the test data set. The evaluation of the model performance is based on their achieved OA, SEN, and SPEC for each class prediction. In addition, the confusion matrix can be generated and evaluated, providing a more detailed view on the prediction performance as it directly shows the amount of correct and incorrect classifications of the test data samples.

Table 5.3 gives an overview of the prediction performance when classifying error-free sound fields, aliasing, noise, and positioning offsets. Results show that the estimates based on the training data were quite accurate, with some models even performing better than expected. Overall, the SGB model performs best with OA = 0.95. All REF and aliasing signals are correctly classified which is indicated by SEN = 1.00 and SPEC = 1.00. For noise and position errors, the model shows slightly lower accuracy with sensitivities of 0.99 and 0.86, and specificity values of 0.99 and 0.98, respectively. In addition, SGB also achieves high Kappa = 0.94 and a NIR = 0.22, indicating that no class imbalances are present on which the model could over-fit. Also, some other models perform extraordinarily well such as ANN, FDA, RF, and BT, with the first three providing accuracies of OA = 0.94 and BT achieving 0.93. They correctly classify error-free sound fields and still show high SEN and SPEC values for aliasing, noise and microphone position errors. The lowest performance is achieved with PLS, providing OA = 0.84. All remaining models still perform well with OA in the range of 0.87–0.91.

Table 5.3: Error classification performances in free-field environments measured with OA, SEN, and SPEC.

Model	OA	REF		Aliasing		Noise		Pos. error	
		SEN	SPEC	SEN	SPEC	SEN	SPEC	SEN	SPEC
LDA	0.89	1.00	1.00	1.00	1.00	0.83	0.99	0.76	0.96
PLS	0.84	1.00	1.00	0.98	1.00	0.83	0.98	0.64	0.94
RDA	0.87	1.00	1.00	1.00	1.00	0.86	0.98	0.73	0.94
ANN	0.94	1.00	1.00	1.00	1.00	0.90	0.99	0.86	0.99
FDA	0.94	1.00	1.00	1.00	1.00	0.87	0.99	0.90	0.97
SVM$_{lin}$	0.87	1.00	1.00	1.00	1.00	0.90	0.97	0.73	0.95
SVM$_{rad}$	0.91	1.00	1.00	1.00	0.99	0.91	0.98	0.86	0.95
KNN	0.89	1.00	1.00	1.00	1.00	0.81	0.99	0.80	0.95
BT	0.93	1.00	1.00	0.99	0.99	0.91	1.00	0.91	0.97
RF	0.94	1.00	1.00	0.99	0.99	0.91	1.00	0.91	0.97
SGB	0.95	1.00	1.00	1.00	1.00	0.99	0.99	0.86	0.98

For a more detailed view on the classification performance, the confusion matrix is exemplarily shown in Table 5.4 for SGB, the best performing model.

Table 5.4: Confusion matrix for error classification with SGB, showing observed and predicted error responses (using all samples available).

	Observed			
Predicted	REF	Aliasing	Noise	Pos. Error
REF	53	0	0	0
Aliasing	0	73	0	0
Noise	0	0	69	1
Pos. Error	0	0	1	58

The confusion matrix gives an overview of the amount of correct and wrong classifications of all 255 samples from the test data set (20 % of all samples). Correct predictions are located on the matrix diagonal. Results show that REF and spatial aliasing are correctly classified for all cases. The amount of correct noise classifications is also excellent, with only one sample being misclassified as an offset error, and vice versa. The following predictor importance analysis shows which MOVs contributes how strong to the classification performance.

Predictor importance

As stated in [141], evaluating the model-dependent predictor importances is likely to be more reliable than statistically analyzing the predictor-outcome relationship if an effective model was built. In the following, a model-dependent predictor analysis is exemplarily conducted for the best performing model, the SGB. Relating to the MOVs in Table 3.2, the overall predictor importances (PI) are given in Figure 5.5, exemplarily showing the 20 most important predictors. Note that importance levels are normalized to a scale from 0 to 100.

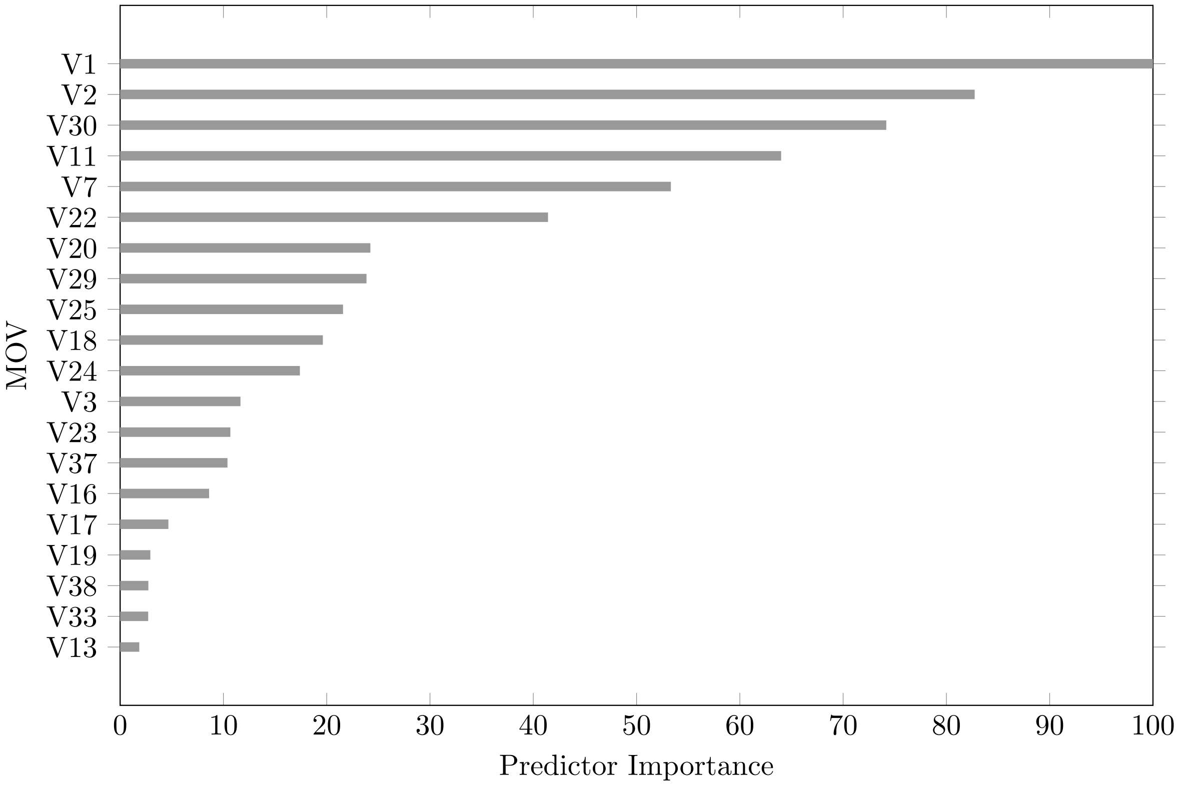


Figure 5.5: Top-twenty predictor importances for error classification using SGB.

Specifically, the modulation differences *winModDiff1* and *avgModDiff1* (V1 and V2) show highest contributions of PI > 80, followed by V30 and V11, the ILD measure $ildMeanFMeanT_{bas}$ and the harmonic structure of error *harmStrucOfErr*, respectively, both still providing contributions above 60. In addition, V7 and V22, the total NMR *totalNmr* and the ITD measure $itdMeanFMeanT_{adv}$, also contribute to the model performance with importances above 40. Using only these six predictors slightly reduces the accuracy of SGB, but also its complexity. Compared to 150 trees and depth = 3 using all predictors, SGB grows only 100 trees using the reduced predictor set down to depth = 1, thereby still achieving an accuracy of 0.91.

This information on predictor importance, however, is only based on the presented multi-class problem. Therefore, no error-specific information can be derived indicating which predictor contributes how strong to the class prediction. In order to derive this information, the data is divided into two categories, introducing a two-class problem with two events: *error* and *non-error*. To account for class imbalances, the amount of samples for *non-error* events is adjusted accordingly. For all errors, again, SGB is used for classification and the respective predictor importance is assessed. Table 5.5 shows the relative importances for each error and the related predictors contributing with PI > 5 while predictors with smaller contributions are discarded from further analysis.

Table 5.5: Relative predictor importances (PI) in ascending order for error classification with SGB. Respective MOVs are shown in Table 3.2

Aliasing		Noise		Pos. error	
MOV	PI	MOV	PI	MOV	PI
V24	100	V1	100	V22	100
V2	82	V2	82	V11	84
V26	26	V11	70	V18	61
V6	6	V37	51	V3	32
-	-	V30	47	V23	32
-	-	V22	45	V24	27
-	-	V7	28	V20	27
-	-	V15	27	V16	27
-	-	V20	22	V7	16
-	-	V25	19	V12	11
-	-	V24	18	V1	11
-	-	V18	17	V13	11
-	-	V29	12	V2	10
-	-	V16	11	V19	10
-	-	V23	7	V25	8

In the table, the predictors are listed in ascending order of importance. Four MOVs mainly account for spatial aliasing: two IACC measures *iaccMaxFMeanT*$_{adv}$ and *iaccMeanFMeanT*$_{adv}$ as well as the average modulation difference *avgModDiff1* and the bandwidth measure for SUT *BandwidthSut*. For noise classification, 15 predictors contribute to SGB performance with relative importances above 5 while the five highest-ranking predictors show contributions above 40, including two measures for modulation differences, namely *winModDiff1* and *avgModDiff1*, as well as the measure for hamonic structure of error *harmStrucOfErr*, the IACC measure *iaccMaxFMaxT*$_{bas}$, and the ILD measure *ildMeanFMeanT*$_{bas}$. Only three MOVs contribute to positioning errors with values greater than 40 such as the ITD measure *itdMeanFMeanT*$_{adv}$, the measure harmonic structure of error *harmStrucOfErr*, and the ILD measure *ildMeanFMeanT*$_{adv}$.

Summary and discussion

In sum, identifying and classifying system errors worked well for all error types regardless of the chosen model. Overall, it can be stated that boosted trees and nonlinear models, e.g., SGB and ANN, provided high classification performances with accuracies above 0.90. They are therefore recommended for error classification.

As indicated earlier in Table 5.3, the classification performance of all models strongly depended on the individual model properties and the respective characteristics of the system error under investigation. Results showed that the discrimination between error-free and erroneous sound fields was easy for all models tested, and that the identification of spatial aliasing was an easy task for most of them. The classification of measurement noise and microphone offset errors achieved slightly lower accuracies, although for both cases, the performance of all models was still above OA = 0.80. In practice, however, the right model choice might also be a question of interpretability and efficiency in terms of real-time applicability.

Following prediction, a predictor importance analysis was conducted, showing that mainly four MOVs contributed strongly to spatial aliasing classification. Fifteen MOVs were needed to discriminate between noise and positioning errors. It was shown that high prediction performance could also be achieved using a reduced sets of predictors, comprising only the top-ranking features.

5.1.2 Regression in Free-Field Environments

Although classification models can confidently identify system errors in spherical microphone array auralizations, they do not provide information on the strength of an error. This, however, might be an important information in practical situations, if one needs to know whether an error might be acceptable or not. Regression models can achieve this task. They are evaluated next.

In this section, a number of predictive models are applied for regression analysis in order to estimate the strength of system errors in spherical microphone arrays. These models are used to estimate the ratings of Experiment I for all three errors, i.e., spatial aliasing, measurement noise, and microphone positioning errors. Basically, the modeling process follows the same procedure as for classification: In a first step, all models are trained and tested against the training data set to find optimal tuning parameters based on the estimated accuracies. In a second step, the developed models are tested against the test data set and their performance evaluated.

Only a limited number of samples is at hand for regression as the only observations available to train the models are the perceptual data from Experiment I. For each error, 60 samples can be used for modeling: The ten test signals from Table 4.1 with six error conditions each. Again, 80 % of the data is used for model training and the remaining 20 % for model testing. Beforehand, and depending on the estimated model performance, the data is preprocessed according to the methods described in Section 3.3.3. For resampling, a 10-fold cross-validation is conducted and repeated 50 times.

According to the estimates obtained from model building, many models seem suitable for error regression. An overview of their performances is given in Figure 5.6, exemplarily showing the estimates for spatial aliasing regression for all models as box plots for R^2 in the right, and RMSEs in the left plot.

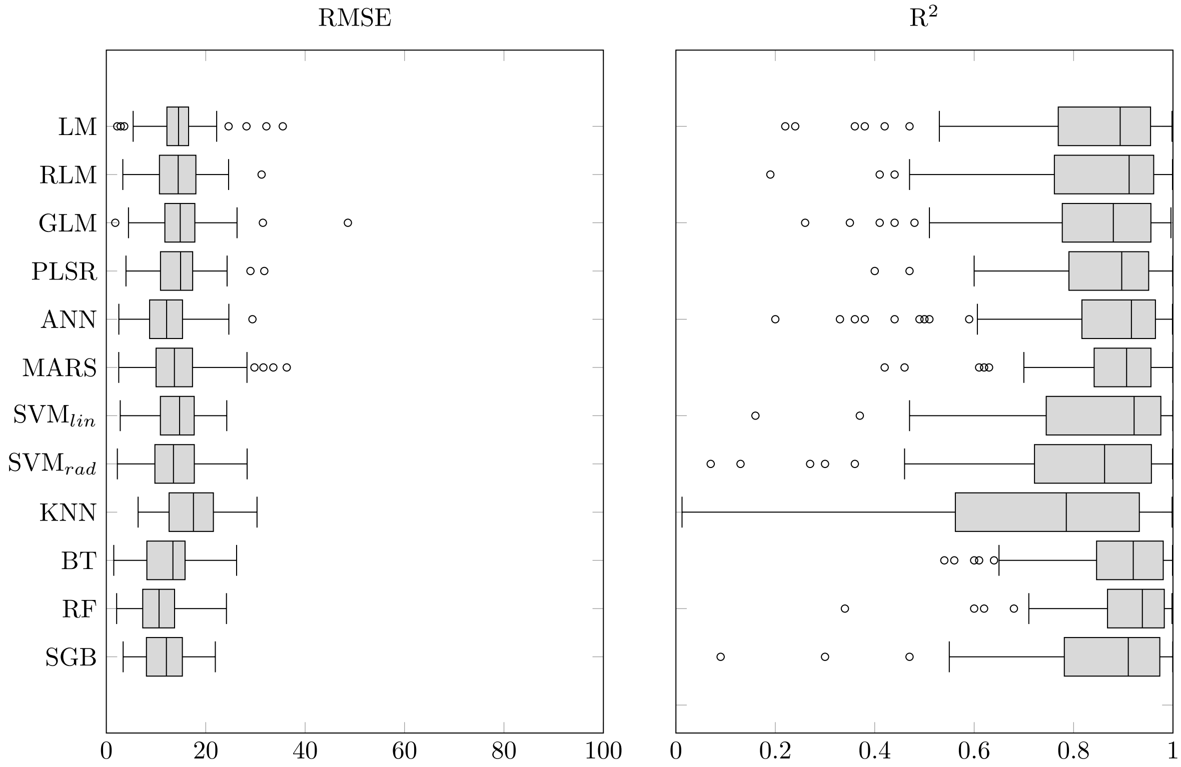


Figure 5.6: RMSE and R^2 values of resampled model performances, estimating spatial aliasing PDGs of Experiment I.

These plots show the variances over the achieved prediction performances after resampling. Consider for example the regression results for KNN in Figure 5.6: Although RMSE estimates are in acceptable limits, the high variances in R^2 would speak against the usage of this model. RF and MARS, on the other side, show the lowest variances in R^2. In the following analysis, only the best performing models in terms of R^2 are evaluated[13]. Information on all other models can be found in the respective Tables C.8, C.9, and C.10 in Appendix C. These tables give an overview of R^2 and RMSE values after resampling, provide performance estimates from model training and tuning, and show the corresponding tuning parameter settings and applied preprocessing steps.

Based on these performance estimates, the following models seem suitable for error regression: RF is expected to perform best estimating spatial aliasing as it provides accuracy and error estimates of $R^2 = 0.90$ and RMSE = 10.55, respectively. Noise and offset errors are best predicted with ANN and SGB, respectively, whereas the first achieves $R^2 = 0.86$ and RMSE = 12.73 and the second $R^2 = 0.76$ and RMSE = 15.80. Comparing the estimated performances over all errors, it becomes clear that most models provide high accuracies for spatial aliasing and noise while performances considerably drop for positioning errors, yielding accuracies in the range of 0.60 to 0.74 with RMSEs around 18. The lowest model performance is expected for BT with $R^2 = 0.47$ and RMSE = 23.42, which is unacceptable in practical applications.

Evaluation based on test data

To see how these models perform on unknown samples, they are applied in the following to estimate the PDGs from the test data set. Table 5.6 gives an overview of the prediction performances in terms of R^2 and RMSE for all models and errors.

Table 5.6: Error prediction performances under free-field conditions for the test data, providing R^2 and RMSE for all models.

Model	Aliasing		Noise		Pos. error	
	R^2	RMSE	R^2	RMSE	R^2	RMSE
LM	0.78	15.08	0.80	10.69	0.66	13.83
RLM	0.86	12.73	0.79	10.98	0.68	14.12
GLM	0.78	15.08	0.80	10.69	0.66	13.83
PLS	0.82	14.22	0.79	10.90	0.68	13.76
ANN	0.82	13.34	0.53	16.76	0.77	11.56
MARS	0.96	8.39	0.42	19.76	0.76	11.55
SVM_{lin}	0.87	12.65	0.79	10.49	0.70	13.57
SVM_{rad}	0.63	20.05	0.44	18.07	0.80	10.52
KNN	0.86	14.59	0.64	14.40	0.68	13.57
BT	0.83	12.34	0.82	11.41	0.50	18.86
RF	0.93	7.79	0.82	10.48	0.79	10.81
SGB	0.95	6.07	0.62	14.90	0.76	11.74

[13] Note that, besides R^2, also RMSE values can be used as a reference to choose a suitable model.

Overall, the ratings for spatial aliasing can be approximated best, followed by noise and positioning offsets. Moreover, it is observed that some models perform better than their estimates suggested, whereas others perform worse. Compared to the other models, MARS and SGB show increased performances, best approximating spatial aliasing with R^2 values of 0.96 and 0.95 and low error scores of 8.39 and 6.07, respectively. The prediction accuracy and residuals are exemplarily depicted in Figure 5.7 for MARS.

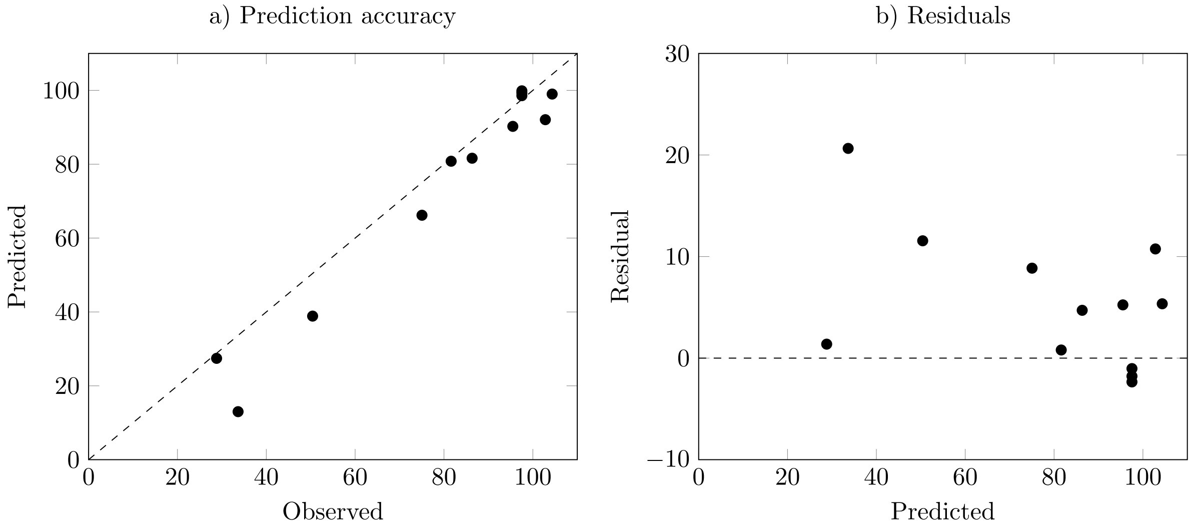


Figure 5.7: Prediction performance evaluation plots using MARS for spatial aliasing regression. Plot a) shows the achieved accuracy, plot b) the residuals.

Plot a) shows the predicted samples over their observed responses, where most samples lie close to the diagonal that indicates perfect prediction, plot b) the residuals.

All other models also show good accuracies in the range of $R^2 = 0.78$–0.87 for aliasing regression. RF, for example, which was expected to perform best, scores slightly lower but still achieves an accuracy of 0.93 with RMSE = 7.79. SVM_{rad} provides the lowest accuracy of $R^2 = 0.63$ with RMSE = 20.05. However, recall that R^2 is the squared correlation coefficient, this model therefore achieves a correlation of nearly 0.80 which should still be high enough in most practical applications. Prediction results for measurement noise are a bit lower. The best performing models are BT, RF, GLM, or LM, showing high accuracies of $R^2 = 0.82$ for the first two, and 0.80 for GLM and LM, with error scores of RMSE = 11.41, 10.48, and 10.69, respectively. Worst performances with $R^2 = 0.42$ and $R^2 = 0.44$ are observed for MARS and SVM_{rad}. For positioning errors, the best results are achieved by SVM_{rad}, scoring with $R^2 = 0.80$ and RMSE = 10.52. RF and ANN also show good accuracies with $R^2 = 0.79$ and 0.77, respectively, while all other models perform worse, providing performances in the range of $R^2 = 0.66$ and 0.76. BT performs worst with $R^2 = 0.50$. Results can also be plotted and compared against the listening test data. This is exemplarily shown in Figure 5.8 for all errors, accordingly using the best-performing model.

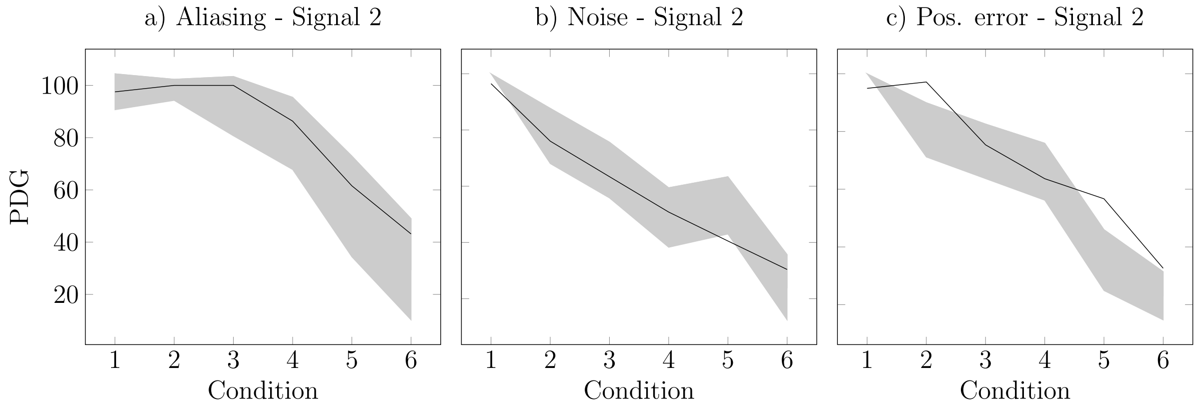


Figure 5.8: Comparison of predicted (solid line) and observed PDGs (gray confidence area) from Experiment I, exemplarily for signal 2 for all errors.

The plots compare the predicted PDGs to the perceptual data exemplarily for signal 2. The predicted PDG is plotted as a solid line over the gray area representing the 95 %-confidence bounds from the listening experiment. Results show that for aliasing and noise in plots a) and b), all predictions lie within this confidence area, whereas small deviations occur for positioning errors, as observable in plot c).

Predictor importance

For this regression analysis, the predictor importance evaluation is twofold: The first is based on model-dependent predictor assessment, for each error evaluating the best performing models: MARS to assess the predictor importance for spatial aliasing regression, BT for noise and SVM_{rad} for positioning error estimation. The second evaluation uses predictor-outcome correlations to identify important predictors. This correlation analysis, however, can only provide directions to draw conclusions since underlying relationships might have been missed. For example, two predictors might show no correlation with the outcome but their interaction does. In this case, the model-dependent predictor importances provide more information on actual relationships in the data. They are given in the following Table 5.7 for all three errors, each exemplarily for its best-perfoming model. The table shows the most important MOVs and their contributions to the model performance in terms of PI.

Table 5.7: Relative predictor importances (PI) for top-ten predictors, in ascending order for error regression with MARS (aliasing), BT (noise), and SVM$_{rad}$ (Pos. error). Respective MOVs are shown in Table 3.2.

Aliasing (MARS)		Noise (BT)		Pos. error (SVM$_{rad}$)	
MOV	PI	MOV	PI	MOV	PI
V1	100	V26	100	V7	100
V29	43	V2	93	V24	91
V12	27	V12	91	V26	83
V21	6	V24	87	V15	81
-	-	V1	83	V5	71
-	-	V7	74	V16	70
-	-	V10	59	V1	69
-	-	V15	59	V10	62
-	-	V3	57	V14	61
-	-	V14	55	V3	58

Note that this table only lists at most the top-ten predictors, comprising features that contribute to the model performce with importance values above 5. MARS mainly uses four predictors for aliasing regression with *winModDiff1* being the most important, contributing with PI = 100 to the model performance. Also *ildMaxFMaxT$_{bas}$* with PI = 43 and *rmsModDiff* with PI = 27 are important MOVs. Only *itdMaxFMaxT$_{adv}$* contributes with PI = 6. To describe all other errors, several more MOVs are important, with all top-ten predictors showing importances above 50. Highest ranking, with PI > 90, are *iaccMeanFMeanT$_{adv}$*, *avgModDiff1*, and *rmsModDiff* for noise error regression using BT as well as *totalNmr* and *iaccMaxFMeanT$_{adv}$* for positioning errors using SVM$_{rad}$.

Finally, the correlations between predictor and outcome are evaluated by choosing a threshold of 0.50 to distinguish between important and less important predictors. For spatial aliasing, seven out of 39 predictors correlate with the outcome, specifically V1, V2, V10, V11, V13, V14, and V18. From these predictors, three show coefficients above 0.70, i.e., *winModDiff1*, *avgModDiff1*, and *avgLinDist*. Four MOVs highly correlate with the outcome of measurement noise regression. These are V1, V2, V12, and V26 which correspond to *winModDiff1*, *avgModDiff1*, *rmsModDiff*, and *iaccMeanFMeanT$_{adv}$*, all providing coefficients above 0.70. Correlations for positioning errors are lower with eight MOVs correlating higher than 0.50: V1, V2, V5, V12, V14, V16, V24, and V26. The highest correlations can be found for *winModDiff1*, *BandwidthRef*, and *iaccMaxFMeanT$_{adv}$* with coefficients of 0.63, 0.62, and 0.68, respectively.

Summary and discussion

All in all, different models performed differently well quantifying the impact of system errors. Depending on the error type under investigation, high regression accuracy was achieved by several models. The best performances for spatial aliasing prediction provided accuracies at $R^2 = 0.95$ or even higher. For measurement noise and offset errors, the highest achieved performance was slightly lower, but still with maximum values at $R^2 = 0.82$ for measurement noise, and accuracies of $R^2 = 0.80$ for offset errors. Classification also worked well with most models scoring in the proximity of OA = 0.90, hence, confidently identifying spatial aliasing, measurement noise, and microphone positioning errors. However, the model's performance decreased from aliasing to noise to offset errors. This might be accounted to shortcomings and methodical errors in the listening experiments, as described in more detail in preceiding Section 4.1. In particular, uncertainties in the perceptual data, i.e., unclear PDG-ratings, are likely to result in low model performances. Specifically for data with high variances and large confidence intervals, such uncertainties create ambiguities when the model tries to map specific predictor values to multiple and also misleading PDG-ratings.

Comparing the predictor-outcome correlations with the model-dependent predictors highlights important MOVs for system error estimation. Specifically, all model-dependent predictors with a high contribution to the model performance (as shown in Table 5.7) were also identified as important predictors for error estimation by the predictor-outcome correlations. Furthermore, the identified predictors were also regarded to be important in the preceding experiment dealing with the classification of system errors in free-field environments. The results from this predictor importance analysis will later be compared to the predictor analysis of the next experiment which deals with system errors under reflective conditions.

5.1.3 Classification in Reverberant Environments

The predictions presented in the last section were based on spherical array simulations under free-field conditions. In this section, the presented modeling approach is applied to the data of Experiment II, addressing system errors in reflective environments. The analysis is similar to the free-field case. A classification approach is applied to categorize system errors and a regression task used to estimate their strength. Subsequently, the respective predictor importances are assessed to identify MOVs relevant for system error description.

The data of Experiment II serves as a basis for classification. All in all, 1680 samples are at hand for predictive modeling. Again, all 70 signals from SQAM are convolved with the BRIRs, each for eight conditions and three error categories. Four classes are predefined: one REF and three error classes. In the analysis, 80 % of the data is used for training and the remaining samples are later used for model testing. Again, for resampling, a 10-fold cross-validation is applied and repeated 50 times.

After model training and tuning, accuracy estimates from resampling indicate that error classification performances are lower under reflective conditions compared to free-field environments (see Tables 5.2 and C.11). While performances in free-field environments were approximately OA = 0.90, most models only achieve accuracies around OA = 0.70 in the presence of room reflections. However, ANN is expected to perform best with OA = 0.81, followed by C50 which achieves OA = 0.79. Applying these models on the test data shows that the performance estimates from resampling were quite accurate, as can be seen in the following Table 5.8.

Table 5.8: Error classification performances in reflective environments, measured with OA, SEN, and SPEC.

Model	OA	REF		Aliasing		Noise		Pos. error	
		SEN	SPEC	SEN	SPEC	SEN	SPEC	SEN	SPEC
LDA	0.72	1.00	1.00	0.99	1.00	0.61	0.89	0.36	0.90
PLS	0.71	1.00	1.00	1.00	1.00	0.57	0.89	0.30	0.90
RDA	0.67	1.00	1.00	1.00	0.99	0.49	0.89	0.25	0.87
ANN	0.84	1.00	1.00	1.00	1.00	0.84	0.97	0.54	0.94
FDA	0.67	1.00	1.00	0.97	0.98	0.48	0.89	0.24	0.88
SVM$_{lin}$	0.65	1.00	1.00	1.00	1.00	0.44	0.87	0.37	0.85
SVM$_{rad}$	0.75	1.00	1.00	0.98	0.99	0.70	0.94	0.33	0.88
KNN	0.71	1.00	1.00	1.00	1.00	0.56	0.93	0.35	0.87
BT	0.77	1.00	1.00	1.00	0.99	0.86	0.91	0.27	0.94
RF	0.77	1.00	1.00	1.00	1.00	0.83	0.95	0.42	0.88
SGB	0.81	1.00	1.00	1.00	1.00	0.81	0.97	0.48	0.91
C50	0.77	1.00	1.00	1.00	0.99	0.85	0.93	0.27	0.92

As indicated by the accuracy estimates after resampling, the best performance of OA = 0.84 is achieved using ANN which categorizes REF and spatial aliasing correctly with SEN = 1.00 and SPEC = 1.00. Performance drops to SEN = 0.84 and SPEC = 0.97 for measurement noise as well as SEN = 0.76 and SPEC = 0.90 for microphone positioning errors. The described trend is similar for all applied models.

Predictor importance

The predictor importances for this classification task are evaluated for ANN which performed best to discriminate the four predefined classes. Thirteen MOVs strongly contribute to the model performance with PI values higher than 50. For this multi-class problem, the five most important predictors are V5, V18, V14, V30, and V11 with contributions of PI = 100, 88, 84, 75, and 75, respectively. To get more detailed information on what predictors contribute to what error, again, each error is analyzed separately in a two-class problem with two events: *Error* and *non-error*. Only the best-performing model ANN is evaluated which relies on MOVs V5, V6, V12, and V33 for aliasing prediction, with contributions of PI = 100 for the first, 97 for the second, 46 for the third, and PI = 17 for the remaining three MOVs. Noise and position errors mainly use five MOVs: V5, V6, V12, V1, and V22 for the first, and V13, V32, V20, V1, and V11 for the latter error.

Summary and discussion

Although the overall performance is lower than in the free-field environment, classification of system errors worked well in reverberant environments. Most models were able to correctly identify and categorize REF and spatial aliasing. However, noise and offset errors were more difficult to discriminate in the presence of room reflections, resulting in lower overall classification performances for some models. The models ANN and SGB scored best while other models such as KNN or PLS still achieved accuracies above OA = 0.70.

5.1.4 Regression in Reverberant Environments

This regression task aims at estimating the strength of system errors under reflective conditions. All data from Experiment II is used for model training and testing. Again, 80 % of the data is used for training and the remaining samples for testing. Note that a direct comparison between measurement errors in dry and reverberant rooms should be handled with care since the two rooms were assessed separately in the perceptual experiments, consequently missing the absolute relation between both.

Tables C.12 to C.14 in the appendix give an overview of the estimated model performances for error regression, showing high prediction accuracies for all errors. Models for spatial aliasing regression perform in the range of $R^2 = 0.75$ and 0.86, providing error scores between 11.14 and 8.72, respectively. Estimates for positioning errors are also high, with most models achieving values around $R^2 = 0.90$ and error scores below 10 while performances for measurement noise attract attention as they are all above 0.95 for all models. PLS and MARS seem to be the best models for aliasing regression, achieving accuracies of $R^2 = 0.86$ and $R^2 = 0.83$, respectively. PLS, MARS, SVM_{rad}, KNN, BT, and RF are all expected to perform with $R^2 = 0.98$ for measurement noise prediction. Positioning errors appear to be best predicted using PLS and RF, both providing accuracy estimates of $R^2 = 0.94$. Although resampling was applied to increase the estimation variance and reduce the chance of over-fitting, most of these values seem overly optimistic in practical applications. They are therefore evaluated against the hold-out test data set to test how these models perform on unknown samples.

The corresponding prediction results are given in the appendix in Table C.15. Overall, prediction performances describe a similar tendency as expected from the estimates after resampling. Spatial aliasing results are in the range of $R^2 = 0.62$–0.90 for all models while prediction performances for noise and offset errors mostly score higher than 0.92. For spatial aliasing regression, SGB performs best with $R^2 = 0.90$ and RMSE = 8.36. Also RF and PLS show good prediction results with $R^2 = 0.87$ and $R^2 = 0.82$, respectively. Noise is best predicted with PLS, achieving an accuracy of $R^2 = 0.98$ with small errors of RMSE = 6.30. Multiple models perform equally well predicting positioning errors with $R^2 = 0.95$ such as LM, RLM, GLM, PLS, ANN, and SVM_{lin}. Among these, RLM provides the lowest error score of RMSE = 7.44.

Predictor importance

By analyzing the predictor importance for regression with ANN, it becomes clear that the model basically uses the same MOVs as in the classification scenario. However, importances slightly vary compared to the classification model. Especially for noise and position error regression, ANN needs more MOVs to achieve high prediction accuracy. For these two error cases, the ten most important predictors all show contributions above 80. Spatial aliasing, however, can be described using only V5, V6, and V12.

Summary and discussion

The results for error regression under reflective conditions are briefly summarized in the following, based on the data derived in Experiment II.

Performances for spatial aliasing prediction showed moderate accuracies around $R^2 = 0.70$. Specifically, decision trees performed best with SGB achieving $R^2 = 0.90$. Predictions for noise and positioning errors, however, appeared to be quite optimistic with values above $R^2 = 0.95$ for most models. This effect might be accounted to methodical issues in the listening experiment, resulting in similar PDGs for most conditions. Consequently, some models are expected to over-fit to these samples, yielding such high prediction performances. Nevertheless, the perceptual effect of noise and positioning errors is strong and results might therefore be correct. However, more test conditions with errors in the medium range should have been included into the listening test design—as it was the case for spatial aliasing regression where the results look more realistic—in order to train the models not only for extreme error degrees. Model performances are therefore expected to be lower when estimating other intermediate error characteristics, at least for microphone noise and positioning errors.

Finally, it should be noted that the same classification and regression analyses were also conducted separately for the dry and reverberant room. Results showed similar accuracies, for some models even higher accuracies and lower error scores, than using the data from both rooms together in a single measurement.

5.1.5 Experiment: Error Comparison between Rooms

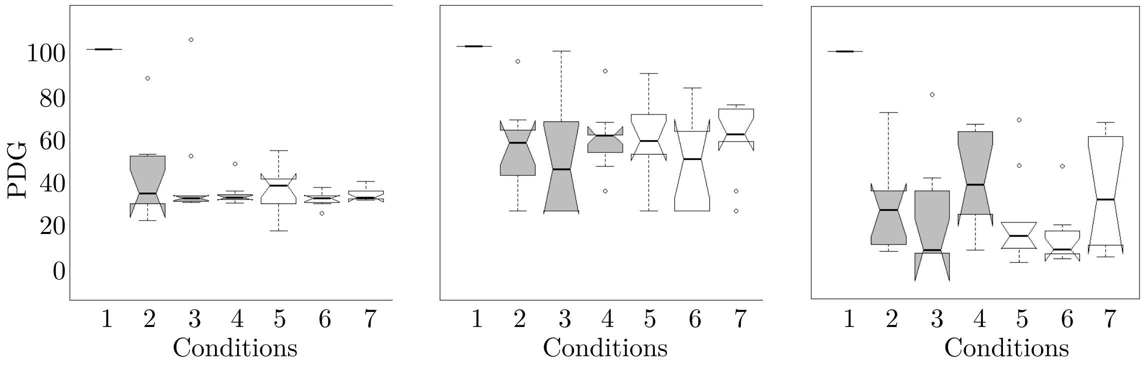
An additional regression analysis is conducted to address the question that could not be answered in Experiment II:

> *Is the perceptual impact of measurement errors depending on the reflection properties of the simulated environment?*

Recall that in Experiment II, errors were assessed separately for the dry and the reverberant room. Consequently, a direct relation between error perception in either room could not be established. In order to answer that question, all models in this analysis are trained and tuned based on the free-field data of Experiment I, and subsequently applied to predict the PDGs of Experiment II. Results are expected to allow statements on the perceptual differences between system errors in dry and reverberant environments. Although absolute PDG scores are expected to be incorrect, the relative relationship between single conditions should still reflect the actual error behavior, assuming that the prediction models only adapt to error characteristics and are uninfluenced by room reflections. Using these models to predict the data in a single experiment would not only establish the desired relationship, but would also allow for a more meaningful predictor importance analysis, yielding suitable MOVs for error description. Moreover, this experiment evaluates the model performances under realistic conditions—in the presence of room reflections.

After model training and tuning (based on a 10-fold cross validation repeated 50 times), only models with the best performance estimates are applied for further evaluation. Using SGB enables aliasing regression with $R^2 = 0.92$ and RMSE = 9.38. The boosted tree RF performs best estimating noise with $R^2 = 0.83$ and RMSE = 13.06. It also provides the highest accuracy to predict positioning errors with $R^2 = 0.77$ and RMSE = 14.66. Performance estimates for all other models are given in Tables C.16 to C.18.

Being trained and tuned with the free-field data, these models are tested by predicting the reflective data comprising the dry and reverberant room. Table C.19 gives an overview of the resulting prediction accuracies and error scores for all applied models. For spatial aliasing regression, it can be seen that the performance of linear models deteriorates with low R^2 around 0.45 and unacceptably high RMSE values, mostly above 65. This is also the case for both SVMs. The trees RF and SGB only achieve R^2 values of 0.46 while BT scores worse with 0.31, all with error scores around 20. However, two nonlinear models, ANN and KNN, achieve moderate to high prediction accuracies of $R^2 = 0.82$ and $R^2 = 0.86$, also providing acceptable error scores of RMSE = 13 and 15, respectively. Both perform well estimating noise and positioning errors with accuracies around 0.70 and above, as well as acceptable RMSEs below 20. Apart from spatial aliasing, also the two trees RF and SGB achieve high accuracies in noise and positioning error prediction, offering similar scores as both nonlinear models. However, in order to estimate the influence of varying room reflections on the perceptual impact of system errors, ANN and KNN are applied. Figure 5.9 displays the results exemplarily for ANN, comparing all errors in the reflective environments.



(a) Aliasing (b) Noise (c) Pos. error

Figure 5.9: Notched box plots for all signals, representing error predictions in reverberant environments based on models trained with free-field data. Condition 1 is REF, conditions 2–4 relate to the dry, and 5–7 to the reverberant room.

Aliasing is shown in plot a), noise in plot b), and positioning errors in plot c). Note that the box plots represent the variances of the predicted PDGs over all ten test signals. Condition 1 is REF, conditions 2 to 4 comprise errors in the dry, conditions 5 to 7 in the reverberant room. They are indicated by gray and white box plots, respectively. Corresponding error conditions are 2 and 5, 3 and 6 as well as 4 and 7. For some cases, it can be seen that the predictions slightly differ between two related conditions, like conditions 2 and 5 in plot c), indicating that the reflection properties could have an impact on error perception, at least in dependence of the auralized test signal. Nevertheless, no significant differences were identified over all conditions. Deviations for single test signals might therefore be an effect of modeling inaccuracy. Note that this result was also verified by using KNN, SGB and RF, all showing no significant differences between system errors in rooms with varying reflection properties. However, even if small differences would in fact be present, the question still remains whether these differences are perceivable or not. This was only evaluated informally by the author who also found no audible difference between corresponding error conditions in both rooms. To fully answer this question, the evaluation of perceptual thresholds and JNDs is proposed for future work.

Predictor importance

For modeling, each model uses its own set of predictors to estimate the outcome. It is therefore hard to state which predictors are actually describing an effect of system errors or just, for example, the variances induced by the characteristics of the test signals. Recall that ten signals were used in the regression, 70 in the classification task.

In order to reveal the causal relationship between error and predictor, a final analysis is made in the following: Since ANN performed well for all errors, the following predictor importance analysis is based on this model. Note that ANN uses PCA predictors, so its importance values relate to the respective component. For each component, only MOVs strongly loading on that component are evaluated, i.e., with loadings above 0.30. Four components, PC3, PC4, PC7, and PC6, are determined important for spatial aliasing description with respective importances of 100, 67, 57, and 49. Specifically, the following MOVs strongly contribute to these components: the ITD measures $itdMaxFMaxT_{adv}$, $itdMeanFMaxT_{adv}$ highly load on PC3 while the average modulation difference *avgModDiff2*, the total NMR *totalNmr*, and the measure for average linear distortions *avgLinDist* contribute to PC4, both bandwidth measures *BandwidthRef* and *BandwidthSut* to PC7, and the average modulation difference *avgModDiff2*, the total NMR *totalNmr*, and the average linear distortions *avgLinDist* to PC6. Measurement noise estimation also relies mainly on the four components PC2, PC6, PC3, and PC5 with importances of 100, 74, 68, and 50, respectively. Important MOVs with high loadings on PC2 are the average modulation difference measure *avgModDiff2*, the bandwidth measure for SUT *BandwidthSut*, and the average linear distortion measure *avgLinDist* while the total and the segmental NMR, *totalNmr* and *segmentalNmr*, the average linear distortions *avgLinDist*, and the ITD measure $itdMaxFMaxT_{adv}$ strongly contribute to PC6. PC3 mainly relies on the average modulation difference *avgModDiff2*, both bandwidth measures *BandwidthRef* and *BandwidthSut*, whereas only two MOVs strongly contribute to PC5: the average linear distortion measure *avgLinDist* and the measure for segmental NMR *segmentalNmr*. Positioning errors can be described using only components PC6 and PC2, providing importance values of 100 and 30. As described above, the corresponding MOVs for the first component are the average modulation difference *avgModDiff2*, the total NMR measure *totalNmr*, and the measure for average linear distortions *avgLinDist*; for the latter, the MOVs for average modulation differences *avgModDiff2*, the bandwidth measure *BandwidthSut*, and the measure for average linear distortions *avgLinDist* strongly contribute. All identified predictors are assumed to account only for the actual error characteristics and are independent of reverberation. KNN, in comparison, mainly relies on four predictors: three measures for modulation differences *rmsModDiff*, *avgModDiff1*, and *avgModDiff2* as well as the average distorted block measure *avgDistBlk*.

Summary and discussion

To establish a relationship between system errors in dry and reverberant environments, a further test was conducted. Results showed no significant difference in the reproduction quality across all errors, no matter whether weak or strong room reflections were present. However, small variances were observed for different test signals. More complex, nonlinear models such as ANN or KNN estimated the strength of spatial aliasing, measurement noise, and positioning errors best. Boosted trees also achieved good results for noise and positioning error prediction, but their performance deteriorated when estimating aliasing. All these models showed high correlations to the PDGs of Experiment II.

A predictor importance analysis of ANN revealed a number of MOVs considered important to predict system errors in reflective environments. All identified predictors also contributed strongly to the performances of RF and SGB, and were also found to be important variables when estimating system errors separately in Experiments I and II. This further underlines their importance for error description and prediction.

5.2 Quality Prediction

The previous sections addressed the prediction of PDGs from Experiments I and II, dealing with the influence of system errors in free-field and reflective environments. This section presents the estimation of the reproduction quality in terms of ASW and LEV using the data derived from Experiment III. All ten audio signals listed in Table 4.1 are auralized based on array simulations using different sound field orders. Three environments with varying reflection properties are taken into account: free-field sound fields serving as ANC and two simulated shoe-box shaped rooms, one with weak and the other with strong reflection characteristics (see Section 4.3 for details).

The following analysis estimates the PDGs from the listening test in Experiment III based on a regression approach and briefly discusses the corresponding predictor importances. Note that a classification task yielded unsatisfying results because ratings of ASW and LEV are too similar, with the outcome that all models fitted on the same characteristics in the data, consequently missing the link between ASW and LEV. The classification results are therefore discarded from this analysis.

5.2.1 Regression Analysis for ASW and LEV Prediction

Seventy samples are available for model building, comprising ten test signals for each of the seven conditions. The data is split into a training and a test set with 80 % being used for model training and 20 % for testing. A 10-fold cross validation is conducted and repeated 50 times to generalize the model performance. Prediction accuracy estimates and preprocessing steps are shown in Table C.20 for ASW, and in Table C.21 for LEV.

Most models are expected to provide high prediction accuracies in the range of $\mathrm{R}^2 = 0.86$ and high, for ASW as well as for LEV predictions. Especially the boosted trees perform well with SGB achieving $\mathrm{R}^2 = 0.95$ for ASW and LEV, RF reaching $\mathrm{R}^2 = 0.95$ for ASW and 0.93 for LEV while BT predicts both with $\mathrm{R}^2 = 0.92$. PLS, ANN, and MARS also show high accuracy estimates above 0.91.

These performance estimates prove to be quite accurate when testing these models against the test data. Table 5.9 shows the achieved accuracies. The models RF, SGB, ANN, and PLS perform best, all scoring with values above 90 and RMSEs below 10. In particular, RF provides an accuracy of $\mathrm{R}^2 = 0.95$ for both ASW and LEV predictions with small errors such as 6.20 for ASW and 7.24 for LEV regression. The worst performance can be observed using SVM_{rad} which, however, still achieves moderate accuracies of $\mathrm{R}^2 = 0.69$ for ASW and $\mathrm{R}^2 = 0.67$ for LEV regression.

Table 5.9: ASW and LEV prediction performances in terms of R^2 and RMSE for all models, predicting the samples from the test data set.

Model	ASW		LEV	
	R^2	RMSE	R^2	RMSE
LM	0.85	11.43	0.86	10.74
RLM	0.87	10.56	0.88	10.38
GLM	0.85	11.43	0.86	10.74
PLS	0.94	7.88	0.94	7.03
ANN	0.91	7.98	0.93	8.66
MARS	0.86	11.65	0.92	8.48
SVM_{lin}	0.87	10.63	0.90	10.41
SVM_{rad}	0.69	15.10	0.67	16.03
KNN	0.85	11.03	0.87	11.29
BT	0.90	9.02	0.88	9.97
RF	0.95	6.20	0.95	7.24
SGB	0.94	7.21	0.92	8.40

5.2.2 Predictor Importance

This predictor analysis is twofold: Firstly, a model-independent evaluation is conducted based on a PCA, evaluating predictor correlations with the outcome. Secondly, the predictor importances of the best performing regression models are assessed and compared to the results of the first analysis. Since importances for ASW and LEV predictions are similar, the following analysis applies to both attributes.

The correlation analysis yields a variety of predictor variables highly correlating with the outcome, all with coefficients above 0.70. These MOVs are the IACC measures V12, V24, V26, V27, V36, V38, and V39 as well as the modulation difference *rmsModDiff*. Results from the PCA indicate that five components explain 95 % of the information, with the first component covering nearly 64 %, the second 14 %, the third 9 %, the fourth 6 %,, and the fifth component only adding additional 2 % to the information. Predictors highly contributing to the first components are V12 and V3 with loadings of 0.72 and 0.53, V5 and V6 loading on PC2 with 0.71 and 0.60, and V3, V12, and V15 on PC3 with contributions of 0.55, 0.31, and 0.30, respectively. Also, several other predictors—mostly spatial—contribute to PC3 with loadings around 0.20. Specifically, the ITDs seem important for ASW and LEV regression. As can be seen in Table 3.2, these MOVs are represented by V20, V21, and V23 from the advanced, as well as V32, V33, and V35 from the basic version of PEAQ.

The last analysis step evaluates predictor importances based on the best performing models. This is exemplarily done for SGB, which mainly relies on five predictors strongly contributing to the model performance. These are the IACC measure $iaccMaxFMeanT_{adv}$, the RMS modulation difference *rmsModDiff*, two further IACC measures $iaccMeanFMeanT_{adv}$ and $iaccMaxFMaxT_{adv}$, as well as the measure for average distorted blocks *avgDistBlk* with respective importances of PI = 100, 51, 50, 48, and 25. It should be noted that these predictor variables also highly contribute to the performances of the other models, despite differing slightly in their predictor choice. These predictors, among others, were also identified as important predictors in the model-independent analysis described earlier. However, as stated before, the derived model-dependent predictor importances are more likely to account for the actual underlying relationships in the data than to evaluate only model-independent importances. The identified variables are therefore considered relevant for ASW and LEV description.

5.2.3 Discussion

The presented prediction experiment dealt with the estimation of ASW and LEV based on a regression approach using the data derived in Experiment III in Section 4.3 for modeling. Results showed that ASW and LEV estimation was possible, with most models providing high prediction accuracies and low error scores. Especially decision trees and nonlinear models, like ANN and MARS, performed well in the regression analysis, achieving accuracies above 0.90 and RMSEs below 10.

The predictor importance analysis indicated that ASW and LEV estimations were mainly based on the evaluation of IACCs and modulation differences. At least regarding the importance of IACCs as suitable predictors, these results confirm the descriptions in the literature on ASW and LEV perception and prediction, for instance in [137]. However, no distinction between ASW and LEV could be achieved since their ratings in the listening test are too similar. As indicated earlier, these similarities yielded all models fitting to the same characteristics in the data, consequently missing the underlying relationships that is important to distinguish between ASW and LEV. This may be due to shortcomings in the listening test, like the lack of head-tracking. This issue was discussed in more detail in the description of Experiment III. In order to overcome this weakness, a listening test is therefore proposed for future work, specifically focusing on the discrimination between ASW and LEV.

5.3 Prediction of Microphone Array Configurations and Sound Field Characteristics

This section presents some additional prediction approaches not directly related to the perceptual analyses of Experiments I–III. Despite estimating the perceptual impact of measurement errors as well as ASW and LEV, further information can also be extracted from the auralized sound field data. To reveal this information which might be of interest in practical applications, two classification experiments are carried out in the following: the first experiment aims at identifying and categorizing the reflection properties of the auralized environment while the second experiment estimates the employed array configuration in terms of the order used for sound field decomposition.

Both classification experiments are based on all MOVs derived from Experiment III, which serve as predictor variables. In addition to the already available 490 samples comprising all 70 test signals from SQAM for each of the seven experimental conditions, also order $N = 5$ and $N = 8$ free-field sound fields are included in the experiment to account for possible class imbalances. Consequently, 630 samples are available for predictive modeling, including three different sound field orders and three reflective environments, i.e., free-field, a dry, and a reverberant room (see Table 4.4 for details). The classification task follows the same procedure as presented above: Firstly, after adequate preprocessing, the data is split into a training and a test data set for further modeling, with 80 % of the data used for training and 20 % for model testing. All models are trained and tuned based on a 10-fold cross validation which is repeated 50 times in order to provide stable performance estimates. Secondly, all models are tested against the hold-out test set to evaluate how they would perform on unknown samples. Prior to the discussion, again, a predictor importance analysis is conducted.

5.3.1 Classification of Microphone Array Configurations

This experiment aims at the identification and categorization of the sound field order employed. For classification, three classes are being predefined: The first represents order $N = 3$ sound fields, the second order $N = 5$, and the third $N = 8$ sound fields. Note that no class imbalances are present in this experiment because samples are equally distributed within all three classes. Table C.22 in the Appendix provides an overview of the estimated classification performances, model tuning parameters, and all applied preprocessing steps.

According to these performance estimates, highest classification accuracies can be achieved with SGB which provides an accuracy of OA = 0.85. Also, RF, BT, and LDA perform well with OA = 0.82 for RF and OA = 0.81 for BT and LDA. All other models still achieve accuracies in the range of OA = 0.66–0.80. In the following, these models are applied to predict the test data set. Results for sound field order classification are shown in Table 5.10 in terms of OA, SEN, and SPEC.

Table 5.10: Sound field order classification performances for all models, measured with OA, SEN, and SPEC.

Model	OA	$N = 3$		$N = 5$		$N = 8$	
		SEN	SPEC	SEN	SPEC	SEN	SPEC
LDA	0.82	0.90	0.86	0.80	0.90	0.71	0.96
PLS	0.81	0.90	0.86	0.73	0.91	0.75	0.57
RDA	0.82	0.93	0.81	0.90	0.90	0.71	1.00
ANN	0.84	0.93	0.79	0.83	0.97	0.71	0.97
FDA	0.67	0.73	0.78	0.67	0.79	0.61	0.93
SVM$_{lin}$	0.74	0.83	0.79	0.73	0.84	0.61	0.96
SVM$_{rad}$	0.76	0.90	0.78	0.70	0.91	0.61	0.93
KNN	0.68	0.78	0.72	0.63	0.82	0.57	0.94
BT	0.75	0.93	0.72	0.60	0.94	0.64	0.93
RF	0.78	0.88	0.86	0.80	0.84	0.61	0.96
SGB	0.81	0.85	0.86	0.83	0.91	0.71	0.93

ANN performs best with OA = 0.84, SEN = 0.93, and SPEC = 0.79 for sound field order 3, SEN = 0.83 and SPEC = 0.97 for order 5, and SEN = 0.71 and SPEC = 0.97 for order 8 sound fields. Good performances around 0.80 are also achieved using the linear models LDA and RDA, or employing PLS and SGB for classification. All other models show moderate accuracies ranging from OA = 0.67 to OA = 0.78. The confusion matrix is exemplarily shown in Table 5.11 for ANN, the best performing model.

Table 5.11: Confusion matrix for sound field order classification with ANN.

	Observed		
Predicted	$N = 3$	$N = 5$	$N = 8$
$N = 3$	35	3	5
$N = 5$	3	25	4
$N = 8$	3	2	20

For a total of 100 available samples, 14 misclassifications occur when predicting sound fields of orders 3 and 8 while only 12 samples are wrongly categorized for $N = 5$.

5.3.2 Classification of Reflection Properties

In this classification task, the reflection properties of the simulated environment are addressed, using three predefined classes: *free-field*, *dry*, and *reverberant*. As in the previous experiment addressing the sound field order classification, all classes are equally balanced due to the added free-field data. Table C.23 gives an overview of all applied models, their expected accuracies, tuning parameters, and data preprocessing steps. Their prediction performance estimates suggest that all models achieve high accuracies above 0.93. RF and SGB are expected to classify reflection characteristics best with an estimated accuracy of OA = 0.98. Following this performance estimation, all models are now evaluated against the left-out test data.

The results for the test data prediction are given in Table 5.12, showing SPEC and SEN values for all models classifying the three environments.

Table 5.12: Reverberation classification performances measured with OA, SEN, and SPEC.

Model	OA	Free-field		Dry		Reverberant	
		SEN	SPEC	SEN	SPEC	SEN	SPEC
LDA	0.98	0.93	1.00	1.00	0.96	0.98	1.00
PLS	0.98	0.93	1.00	1.00	0.96	0.98	1.00
RDA	0.95	0.93	1.00	1.00	0.91	0.90	1.00
ANN	0.97	0.93	1.00	1.00	0.95	0.95	1.00
FDA	0.95	0.93	0.99	0.98	0.93	0.93	1.00
SVM$_{lin}$	0.96	0.93	1.00	1.00	0.93	0.93	1.00
SVM$_{rad}$	0.98	1.00	0.99	1.00	0.98	0.95	1.00
KNN	0.95	0.93	1.00	1.00	0.91	0.90	1.00
BT	0.98	0.93	1.00	0.98	0.98	1.00	0.98
RF	0.99	0.93	1.00	1.00	0.98	1.00	1.00
SGB	0.98	0.93	1.00	1.00	0.96	0.98	1.00

As already indicated by the performance estimation during model training, RF achieves the highest accuracy with OA = 0.99, accurately classifying free-field sound fields with SEN = 0.93 and SPEC = 1.00. The identification of dry environments works also well with the same model, achieving accuracies of SEN = 1.00 and SPEC = 0.98. Strong reflective environments, i.e., the reverberant room, can be discriminated from all other classes without a single misclassification as indicated by SEN = 1.00 and SPEC = 1.00. Although RF achieves the highest accuracy with OA = 0.99, also all other models perform very well with prediction performances all above OA = 0.95.

5.3.3 Predictor Importance

The predictor importances for both preceding experiments are evaluated next, taking only the model-dependent predictors into account. Specifically ANN and RF, the best-performing models, are evaluated to identify predictors important for the classification of the employed sound field order and the reverberation characteristics. The first model relates to the used array configuration and the second to the conditions of the simulated environmental in terms of reverberation.

ANN mainly relies on five predictors to identify the used sound field order. Namely, *totalNmr*, $ildMeanFMeanT_{bas}$, *avgModDiff1*, *avgLinDist*, and *BandwidthRef* contribute to the model performance with corresponding importances of PI = 100, 66, 57, 47, and 41. For reverberation detection, RF basically uses two predictors, $iaccMaxFMeanT_{adv}$ and *totalNmr*, mainly contributing to the classification performance with importances of PI = 100 and PI = 25, respectively. The fact that most of the information appears to be stored in the IACCs is plausible since, as explained earlier, they are directly related to the perception of spaciousness.

5.3.4 Discussion

All presented prediction models were applied to confidently categorize technical characteristics of the array and the simulated environment. This was evaluated based on two classification approaches: In a first experiment, predictive modeling was applied to classify different array configurations in terms of varying sound field orders. Results showed that most models performed well for this task, with the highest accuracy, OA = 0.84, being achieved using ANN for classification. The forest RF performed best in the second experiment to discriminate the data into free-field, dry, and reverberant environments. Specifically in this task, the model achieved a remarkably high prediction performance of OA = 0.99. Although such high accuracies might commonly be the result of over-fitting, this is not assumed in this experiment because all three environments differ significantly in their reflective properties. Consequently, it is not surprising that all three classes were clearly separable. Otherwise, if the reflection properties would only slightly vary between conditions, then some models are assumed to perform worse, whereas in this case, a more careful model training and tuning is expected to increase their classification accuracy.

6 Conclusions

The scope of this thesis was on quality assessment of binaural spherical microphone array auralizations for room simulation applications and, in particular, on predictive modeling with the aim to estimate the reproduction quality. The investigations carried out were basically divided into two parts: In the first part, a comprehensive perceptual analysis of the measurement system was conducted, addressing the influence of the array configuration and system errors on the reproduction quality. The second part comprised quality predictions based on the combination of the perceptual data and the output of an auditory model, with the goal to estimate the results of the listening experiments in regression analyses. In addition, various classification tasks were conducted to categorize system errors, the array configuration in terms of the sound field decomposition order, and the reflection properties of the simulated room.

Results show that it is possible to develop well-performing prediction models for the assessment of spherical microphone array auralizations, with most models achieving moderate to high prediction accuracies and low error scores, at least for the conditions evaluated in the presented experiments. In the following, the experimental results are described and discussed in more detail.

6.1 Perceptual Evaluation

The perceptual evaluations in this thesis comprised three listening experiments: The first two dealt with the assessment of measurement errors based on quantitative analyses, whereas the third evaluated the auralization quality using two descriptive attributes from concert hall acoustics: ASW and LEV. All presented VAEs were based on spherical microphone arrays simulated under free-field conditions and in two shoe-box shaped rooms with varying reflection properties. In the simulations, three kinds of system errors were addressed: spatial aliasing, measurement noise, and normally distributed microphone positioning errors. Ten different signals taken from the SQAM data base served as test material in the auralizations.

For all errors, the first two experiments showed that successive error increments led to a stepwise decrease of the reproduction quality, with small error characteristics still being considered acceptable. In addition, results showed that the impact of system errors strongly depended on the characteristics of the presented audio material. In general, no significant influence of the reverberation properties of the simulated environment on the perceived strength of each error could be proven, although for some signals a slight difference was observed.

The third experiment assessed the auralization quality in a descriptive analysis based on ASW and LEV. Again, ten test signals were presented in three simulated environments with varying reflection properties: under free-field conditions as well as in a dry and a reverberant shoe-box shaped room. Assessors had to rate the influence of three different array configurations in terms of varying sound field orders, particularly comparing orders $N = 3$, 5, and 8. In accordance with the literature, results show that ASW and LEV ratings were higher in the reverberant room and that an increase of the array order yielded a significant increase in ASW and LEV. This, at least, was the case for all signals in the strong reverberating room. In the dry room, however, a contrasting behavior was observed: An order increment from $N = 5$ to $N = 8$ resulted in reduced ASW and LEV, although the opposite was observed from order $N = 3$ to $N = 5$, as would be expected from theory where higher orders are associated with higher quality [8]. However, recent experiments in [28] showed that, especially when employing Lebedev sampling (like in the presented experiments), it may occur that specific higher orders result in lower quality than lower orders. In the presented experiments, this effect was significant for some, and tendential for all other signals, and it was only observed in the dry room but not in the strong reverberating environment. Moreover, a similar effect was also observed in [194]: Specifically designed to predict ASW and LEV in binaural signals, the model developed in [137] was applied to estimate the ratings of Experiment III. Results show high correlations with the listening test data, also accurately predicting the discussed effect. Since the applied prediction model was developed and trained with different data, it can be stated that the observed behavior is not only an artifact, for example due to a methodical error in the test design, but actually related to the physical characteristics stored in the binaural signals. In conclusion it is stressed that the perception of ASW and LEV in spherical array auralizations depends on the interaction between the test signal, the employed sound field order, and the reflection properties of the simulated environment.

6.2 Predictive Modeling

The second part of this thesis dealt with predictive modeling, using various models for regression and classification, like linear models, nonlinear models, and also different kinds of decision trees such as random forests or boosted trees. The modeling experiment was divided into three parts: In the first, the goal was to predict the results of the listening experiments in a regression approach estimating the strength of various error characteristics, whereas the second task aimed at predicting the assessors' responses in terms of ASW and LEV. In these tasks, the data collected in the listening experiments served as observations and the output of the PEAQ model as predictors, hence enabling quality prediction. The third part dealt with the identification and categorization of system errors, array configurations, and the reflection properties of the auralized room based on classification experiments.

Results show that the impact of system errors and the auralization quality in terms of ASW and LEV could confidently be predicted with most models. They also performed well in classifying the employed array configuration and the reflection properties of the simulated room. In practice, the right model choice would be a trade-off between robustness, interpretability, and complexity in terms of real-time applicability. For example, in academic research, high prediction accuracy might be the goal while model complexity plays a minor role, whereas less complex models are recommended when real-time applicability is desired, which, on the downside, might provide lower prediction performance. In this context, the results from this thesis provide a guideline for basic model building and application, enabling further analyses of VAEs based on spherical microphone arrays without the need to conduct time-consuming listening experiments. However, reasonable results are only expected if the same experimental conditions are met. For example, it is possible to test further sound field orders and their influence on ASW and LEV perception specifically in the three evaluated environments, using the same ten test signals. Under these conditions, the models are expected to provide realistic prediction results, although a perceptual verification, especially when predicting unknown data, is always recommended—at least by the practicing engineer. Furthermore, a regression analysis was conducted to compare the influence of system errors in different reflective environments, as this question was not addressed in the listening experiments. To answer this question, all models were trained only with errors under free-field conditions and subsequently applied to predict their impact in reflective environments. Only nonlinear and tree-based models achieved reasonable performances as they adapted on the underlying error characteristics in the data, mostly uninfluenced by the reflection properties of the simulated environment. Results show that no significant difference in quality was found when presenting a specific error with a certain degree in environments with varying reflection properties.

For each modeling task, predictor importances were evaluated in order to identify predictors highly contributing to the model performance, thus describing the effect under investigation. In this regard, the auditory model used in PEAQ provides suitable measures to quantify the impact of system errors and sound field characteristics in a way to confidently predict ASW and LEV, to identify and categorize system errors, and to classify properties of the array or the simulated environment.

6.3 Limitations

This section discusses limitations of the presented analysis. Firstly, not all required quality criteria for binaural synthesis could be met in the perceptual experiments such as a dynamic binaural synthesis. This is known to increase the reproduction quality which could have improved the rating accuracy of the assessors, therefore the data basis for modeling. However, in the scope of the three conducted listening experiments, results are still regarded as meaningful: In the error analysis, spatial impression played a minor role, specifically in the presence of annoying artifacts, as was shown in [153]. Although it is assumed that head-tracking could have improved the perception of ASW and LEV, results indicate that they were still confidently assessable using a static synthesis. In addition, evaluating head-tracked binaural signals with an auditory model would have gone beyond the scope of this thesis and is proposed for future work. Secondly, design flaws in the perceptual experiments, specifically in Experiment II, led to ambiguities and inaccuracies in the data which also affected the predictions. Here, a more concise listening test design could have increased the model performance, like training assessors on error-specific sound characteristics, which is likely to result in more meaningful ratings. Thirdly, only a limited amount of samples was available for model building and evaluation. Although resampling was applied to generalize the model performance, and a sample set was hold out for model testing, it is recommended to take more data into account when developing quality prediction models for spherical microphone arrays in practical applications. This includes the amount of test signals, the resolution of the presented test conditions, i.e., error characteristics, the amount of sound field decomposition orders, different environments with various reflection properties, also taking other room geometries into account.

Although most models showed moderate to high prediction performances, it has to be stated that all presented results are only valid under their experimental conditions, specifically the analyzed use cases: the ten auralized test signals taken from SQAM, the three simulated environments, the three types of system errors as well as the specific sound field orders investigated in terms ASW and LEV. Within these conditions, all models are expected to provide reasonable prediction results, in particular when addressing other array configurations. Nevertheless, care has to be taken when applying these models to predict unknown samples. In this case, an accompanying informal perceptual test is always recommended to verify the prediction results. Due to the lack of generalization, this thesis is considered a feasibility study, showing how a quality prediction system for spherical microphone array auralizations could look like.

7 Future Work

The previous section discussed limitations of the experiments presented in this thesis, consequently providing first clues for future research. Overall, more use cases in terms of test signals, VAE data, and array configurations need to be evaluated. Especially, the array configuration is a prominent factor affecting the reproduction quality which needs more addressing. Although it is known that a higher sound field order leads to an increased spatial response of the array, its impact on quality is not yet clear. As indicated by the results of Experiment III, an increased sound field order does not necessarily lead to an increased quality, at least in terms of ASW and LEV. In addition, the author (and some colleagues at TU Ilmenau and Fraunhofer IDMT) stumbled upon this issue during their work on quality assessments of spherical microphone array auralizations. Moreover, recently presented results in [28] and [193] underline the importance of this problem. In this regard, the developed prediction models can provide hints on what sound field orders should be analyzed in a more detailed perceptual evaluation.

In the context of spherical microphone arrays for auralization applications, some further ideas for future research are proposed in the following. For example when evaluating quality, in addition to a *perfect* binaural reproduction (in a technical sense), accompanying visual cues should be added to the reproduction. This could help to identify auditory thresholds which are important clues to define the minimum accuracy needed by the technical system, i.e., the microphone array. In this regard, the assessment of *authenticity* is applicable, while *plausibility* should be used as a quality metric when no visual clues are included. Moreover, advanced quality assessment methods, like OPQ, could be applied to address the auralization quality in more detail, thereby also increasing the prediction accuracy. In addition, JNDs of system errors should also be addressed. In addition to spatial aliasing, microphone noise, and positioning offsets, the quality of the auralization is also influenced by the (frequency dependent) directivity characteristic of the used microphones. Moreover, the application of other auditory models could provide further suitable predictors which might lead to a more detailed information of the sound field characteristics and, consequently, further increase the prediction accuracy. In this context, also more advanced prediction models should be applied, like deep neural networks, as they are likely to improve the prediction performance. And finally, of course, the impact of the measurement system on the auralization quality should be addressed in future research, when for example using loudspeaker-based systems for reproduction such as WFS or HOA.

However, in the author's opinion, sound field representations in spherical coordinates, and specifically their binaural auralization, is the most promising approach for realistic 3D audio reproduction in VR applications—at least for the time being. In order to record content for such auralizations, spherical microphone arrays are the systems of choice due to their flexibility in terms of sound field manipulation abilities based on digital processing and their independence of the reproduction system. Although spherical arrays have been investigated and described very well from an analytical point of view, and recent publications contributed to the state of research regarding their reproduction quality, several questions are still unanswered and need further addressing. In this regard, this thesis contributes to the common literature with some insight on the perceptual influence of system errors and the array configuration. The developed prediction models prove to be suitable tools for automated quality analysis when evaluating spherical microphone arrays for auralization applications.

List of Figures

1.1 Virtual acoustic environments schematic 3

1.2 Some spherical microphone arrays 6

1.3 Outline of this thesis as workflow schematic 9

3.1 Set of spherical harmonics of orders $n = 0 \ldots 2$ 21

3.2 Directivity pattern and PWD half resolution 23

3.3 Sampling quadratures 27

3.4 Mode strength of different array configurations 31

3.5 Error contributions E for various array configurations 43

3.6 Processing stages of the human auditory system (after [258]). 46

3.7 Threshold in quiet (after [274]). 47

3.8 Level of the test tone masked by wide band white noise (after [310]). . . 48

3.9 Masking thresholds for pure tones masked by white noise with critical bandwidth 49

3.10 Masking thresholds for pure tones masked by band-limited noise 50

3.11 Masking thresholds for pure tones masked by pure tones 50

3.12 Principle of temporal masking 51

3.13 Relation of critical bands to cochlea and frequency (after [310]). 52

3.14 Principles of ITD and ILD 54

3.15 Principle of ASW 56

3.16 Principles of LEV 58

3.17 High-level representation of the model employed in PEAQ. 59

3.18 Schematic of an artificial neural network 67

4.1 Schematic of a notched box plot 81

4.2 PDGs for spatial aliasing in free-field simulations 84

4.3 PDGs for measurement noise in free-field simulations 85

4.4 PDGs for microphone positioning errors in free-field simulations 86

4.5 PDGs for spatial aliasing in simple room geometries 90

4.6 PDGs for measurement noise in simple room geometries 91
4.7 PDGs for microphone positioning errors in simple room geometries . . . 92
4.8 ASW ratings as notched box plots for signals 6 and 7 95
4.9 LEV ratings as notched box plots for signals 6 and 7 97

5.1 MOV correlation matrix showing between-predictor-correlations 103
5.2 Predictor skewness for error classification 104
5.3 Explained variances for PCA components 105
5.4 Parameter tuning example 107
5.5 Predictor importance for error classification using SGB 110
5.6 RMSE and R^2 values of resampled model performances 113
5.7 Accuracy and residuals for aliasing regression with MARS 115
5.8 Comparison of predicted and real PDGs from Experiment I 116
5.9 Results of prediction experiment 123

A.1 Schematic of a room impulse response 177

B.1 Spherical coordinate system 181
B.2 Legendre polynomials $P_n(cos(\theta))$ 184
B.3 Associated Legendre functions 185
B.4 Spherical Bessel and Neumann functions and derivatives 189

D.1 PDGs of spatial aliasing in free-fields 207
D.2 PDGs of measurement noise error in free-field 208
D.3 PDGs of microphone positioning errors in free-fields 209
D.4 PDGs of spatial aliasing in simple room geometries 210
D.5 PDGs of measurement noise in simple room geometries 211
D.6 PDGs of microphone positioning errors in simple room geometries 212
D.7 PDGs for ASW for signals 1, 2, 3, 4, 5, 8, 9, and 10 213
D.8 PDGs for LEV for signals 1, 2, 3, 4, 5, 8, 9, and 10 214

List of Tables

3.1 Array configurations used to illustrate error contributions 42
3.2 MOVs of PEAQ model with respective abbreviation. 63

4.1 Test signals used in perceptual experiments 80
4.2 Experiment I listening test conditions 82
4.3 Experiment II listening test conditions 88
4.4 Experiment III listening test conditions 94

5.1 Overview of predictive models used in the experiments 101
5.2 Model parameters for error classification 108
5.3 Error classification performances 109
5.4 Confusion matrix for error classification with SGB 109
5.5 Error classification predictor importances 111
5.6 Error prediction performances in free-field environments 114
5.7 Error regression predictor importances 117
5.8 Error classification performances in reflective environments 119
5.9 Overview of R^2 and RMSE for ASW and LEV prediction based on test data set for all models 127
5.10 Sound field order classification performances 130
5.11 Confusion matrix for sound field order classification with ANN. 130
5.12 Reverberation classification performances 131

C.1 Frequency dependent scattering coefficients 193
C.2 Frequency dependent absorption coefficients for low reverberation. 193
C.3 Frequency dependent absorption coefficients for high reverberation. . . . 193
C.4 PCA loadings of predictor prescreening for all errors 194
C.5 PCA loadings of predictor prescreening for spatial aliasing 195
C.6 PCA loadings of predictor prescreening for measurement noise 196
C.7 PCA loadings of predictor prescreening for microphone position errors . . 197

C.8 Estimated model performances for spatial aliasing regression in free-field environments . . . 198
C.9 Estimated model performances for measurement noise regression in free-field environments . . . 198
C.10 Estimated model performances for positioning error regression in free-field environments . . . 199
C.11 Estimated model performances for error classification in reverberant environments . . . 199
C.12 Estimated model performances for spatial aliasing regression in reverberant environments . . . 200
C.13 Estimated model performances for noise regression in reverberant environments . . . 200
C.14 Estimated model performances for positioning error regression in reverberant environments . . . 201
C.15 Error prediction performances in reverberant environments . . . 201
C.16 Estimated model performances for aliasing regression in reverberant environments using models trained with free-field data . . . 202
C.17 Estimated model performances for noise error regression in reverberant environments using models trained with free-field data . . . 202
C.18 Estimated model performances for positioning error regression in reverberant environments using models trained with free-field data . . . 203
C.19 Error prediction performances in reverberant environments using models trained with free-field data . . . 203
C.20 Estimated model performances for ASW regression . . . 204
C.21 Estimated model performances for LEV regression . . . 204
C.22 Estimated model performances for sound field order classification . . . 205
C.23 Estimated model performances for room classification . . . 205

List of Acronyms

ADA **ADA**boost

AIC **A**kaie **I**nformation **C**riterion

AITD **A**verage **I**nteraural **T**ime **D**ifference

ANC **ANC**hor Signal

ANN **A**rtificial **N**eural **N**etwork

ASW **A**pparent **S**ource **W**idth

AUC **A**rea **U**nder **C**urve

BI **B**ass **I**ndex

BQI **B**inaural **Q**uality **I**ndex

BRIR **B**inaural **R**oom **I**mpulse **R**esponse

BT **B**oosted **T**ree

DBS **D**ynamic **B**inaural **S**ynthesis

DC **D**irect **C**urrent

DFT **D**iscrete **F**ourier **T**ransform

DFTF **D**iffuse **F**ield **T**ransfer **F**unction

DI **D**irectivity **I**ndex

DIX **D**istortion **I**nde**X**

ERB **E**quivalent **R**ectangular **B**andwidth

FBR **F**ront to **B**ack **R**atio

FCP **F**ree **C**hoice **P**roviling

FDA **F**lexible **D**iscriminant **A**nalysis

FFT **F**ast **F**ourier **T**ransform

FIR **F**inite **I**mpulse **R**esponse

GLM **G**eneralized **L**inear **M**odel

HATS **H**ead **A**nd **T**orso **S**imulator

HOA **H**igher **O**rder **A**mbisonics

HRIR **H**ead **R**elated **I**mpulse **R**esponse

HRTF **H**ead **R**elated **T**ransfer **F**unction

IACF **I**nter**A**ural **C**ross correlation **F**unction

IACC **I**nter**A**ural **C**ross correlation **C**oefficient

ITDG **I**nteraurak **T**ime **D**elay **G**ap

IC **I**nteraural **C**orrelation

IIR **I**nfinite **I**mpulse **R**response

ILD **I**nteraural **L**evel **D**ifference

IN **I**nternal **N**oise

ITD **I**nteraural **T**ime **D**ifference

ITU **I**nternational **T**elecommunication **U**nion

JND **J**ust **N**oticable **D**ifferences

JNLD **J**ust **N**oticable **L**evel **D**ifferences

KNN **K** **N**earest **N**eighbor

LEV **L**istener **E**n**V**elopment

LC **L**ateral **C**omponent

LCMV **L**inearly **C**onstrained **M**inimum **V**ariance

LDA **L**ineear **D**iscriminant **A**nalysis

LG **L**ateral (Hall) **G**ain

LM **L**inear **M**odel

LRR **L**eft to **R**ight **R**atio

MARS **M**ultivariate **A**daptive **R**egression **S**plines

MDA **M**ixture **D**iscriminant **A**nalysis

MIC **M**aximal **I**nformation **C**oefficient

MOS **M**ean **O**pinion **S**core

MOV **M**odel **O**utput**V**ariable

MVDR **M**inimum **V**ariance **D**istortionless **R**esponse

NAH **N**ear-field **A**coustic**H**olography

NIPALS **N**onlinear **I**terative **P**artial **L**east **S**quares

NIR **N**o-**I**nformation **R**ate

NMN **N**oise **M**asking **N**oise

NMT **N**oise **M**asking **T**one

OA **O**verall **A**ccuracy

OASE **O**bjective **A**udio **S**ignal *Eval*uation

OPQ **O**pen **P**rofiling of **Q**uality

PAQM **P**erceptual **A**udio **Q**uality *M*easure

PC **P**rincipal **C**omponent

PCA **P**rincipal **C**omponent *A*nalysis

PDG **P**erceptual **D**ifference *G*rade

PEAQ **P**erceptual **E**valuation of **A**udio **Q**uality

PEAQ-MC **P**erceptual **E**valuation of **A**udio **Q**uality - **M**ulti-**C**hannel

PERCEVAL **Perc**eptual *E*valuation

PLS **P**artial **L**east *S*quares

POM **P**erceptual **O**bjective *M*easure

PWD **P**lane **W**ave *D*ecomposition

PWE **P**lane **W**ave *E*xtrapolation

QESTRAL **Q**uality **E**valuation of **S**patial **T**ransmission and **R**eproduction using an **A**rtificial **L**istener

QDA **Q**uadratic **D**iscriminant **A**nalysis

QoE **Q**uality **o**f **E**xperience

RDA **R**egularized **D**iscriminant **A**nalysis

REF **R**eference signal

RF **R**andom **F**orrest

RIR **R**oom **I**mpulse **R**esponse

RLM **R**obust **L**inear **M**odel

ROC **R**eceiver **O**perating **C**haracteristic

RMSE **R**oot **M**ean **S**quared **E**rror

RT **R**everberation **T**ime

SAQI **S**patial **A**udio **Q**uality **I**nventory

SEACEN **S**imulation and **E**valuation of **AC**oustic **EN**vironments

SEN **SEN**sitivity

SFT **S**pherical **F**ourier **T**ransformation

SGB **S**tochastic **G**radient **B**oost

SH **S**pherical **H**armonic

SHD **S**pherical **H**armonic **D**ecomposition

SHE **S**pherical **H**armonic **E**xtrapolation

SNR **S**ignal to **N**oise **R**atio

SPEC **SPEC**icifity

SR **S**lope **R**ate

SSE **S**um of **S**quared **E**rror

SUT **S**ignal **U**nder **T**est

SVM **S**upport **V**ector **M**achine

VAE **V**irtual **A**coustic **E**nvironment

VR **V**irtual **R**eality

WFS **W**ave **F**ield **S**ynthesis

WNG **W**hite **N**oise **G**ain

Bibliography

[1] ISO Standard 3382-1:2009. *Acoustics – Measurement of room acoustic parameters – Part 1: Performance spaces*. 2009.

[2] ISO Standard 3382-2:2008. *Acoustics – Measurement of room acoustic parameters – Part 2: Reverberation time in ordinary rooms*. 2008.

[3] ISO Standard EN 9000:2005. *Quality management systems – Fundamentals and vocabulary*. 2005.

[4] ISO Standard EN 9001:2008. *Quality management systems Requirements*. 2008.

[5] H. Akaike. "A new look at the statistical model identification." In: *IEEE Transactions on Automatic Control* 19.6 (1974), pp. 716–723. DOI: 10.1109/TAC.1974.1100705.

[6] D. Alon and B. Rafaely. "Beamforming with Optimal Aliasing Cancellation in Spherical Microphone Arrays." In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24.1 (2016), pp. 196–210. DOI: 10.1109/TASLP.2015.2502059.

[7] N.S. Altman. "An introduction to kernel and nearest-neighbor nonparametric regression." In: *The American Statistician* 46.3 (1992), pp. 175–185. DOI: 10.1080/00031305.1992.10475879.

[8] A. Avni, J. Ahrens, M. Geier, S. Spors, H. Wierstorf, and B. Rafaely. "Spatial perception of sound fields recorded by spherical microphone arrays with varying spatial resolution." In: *Journal of the Acoustical Society of America* 133.5 (2013), pp. 2711–2721. DOI: 10.1121/1.4795780.

[9] M. Barron. "Late lateral energy fractions and the envelopment question in concert halls." In: *Applied Acoustics* 62.2 (2001), pp. 185–202. DOI: 10.1016/S0003-682X(00)00055-4.

[10] M. Barron. "Subjective study of British symphony concert hall." In: *Acta Acustica united with Acustica* 66.1 (June 1988), pp. 1–14.

[11] M. Barron. "The subjective effects of first reflections in concert halls – the need for lateral reflections." In: *Journal of Sound and Vibration* 15.4 (1971), pp. 475–494. DOI: 10.1016/0022-460X(71)90406-8.

[12] M. Barron and A. H. Marshall. "Spatial Impression due to early lateral reflections in concert halls: The derivation of a physical measure." In: *Journal of Sound and Vibration* 77.2 (1981), pp. 211–232. DOI: 10.1016/S0022-460X(81)80020-X.

[13] R. Baumgartner, P. Majdak, and B. Laback. "Modeling sound-source localization in sagittal planes for human listeners." In: *Journal of the Acoustical Society of America* 136.2 (2014), pp. 791–802. DOI: 10.1121/1.4887447.

[14] S. Bech and N. Zacharov. *Perceptual Audio Evaluation - Theory, Method and Application*. Chichester: John Wiley & Sons Ltd, 2006. ISBN: 978-0-470-86923-9.

[15] J. Becker and M. Sapp. "Synthetic soundfields for the rating of spatial perceptions." In: *Applied Acoustics* 62 (2001), pp. 217–228. DOI: 10.1016/S0003-682X(00)00057-8.

[16] J. Beerends and J. Stemerdink. "Measuring the quality of audio devices." In: *Proceedings of the 90th AES Convention, preprint No. 3070*. Paris, France, 1991.

[17] J. Beerends and J. Stemerdink. "Perceptual evaluation of speech quality (PESQ), the new ITU standard for end-to-end speech quality assessment, Part II – Psychoacoustic model." In: *Journal of the Audio Engineering Society* 50.10 (2002), pp. 765–778.

[18] D.R. Begault. *3-D Sound for Virtual Reality and Multimedia*. Tech. rep. Moffett Field, CA, USA: NASA - Ames Reserach Center - Technical Memorandum, 2000.

[19] R. Bellman. *Dynamic Programming*. Princeton University Press, 1957. ISBN: 978-0-691-07951-6.

[20] I. Ben Hagai, F. Fazi, and B. Rafaely. "Efficient dual-sphere microphone array design based on generalized sampling theory." In: *Proceedings of the EAA Forum Acusticum*. Aalborg, Denmark, 2011, pp. 2221–2226.

[21] L. Beranek. "The sound strength parameter G and its importance in evaluating and planning the acoustics of halls for music." In: *Journal of the Audio Engineering Society* 129.5 (May 2011), pp. 3020–3026. DOI: 10.1121/1.3573983.

[22] L.L. Beranek. *Music, Acoustics, and Architecture*. New York, USA: John Wiley & Sons, 1962.

[23] J. Berg. “Evaluation of Perceived Spatial Audio Quality.” In: *Journal of Systemics, Cybernetics, and Informatics* 4.2 (2006), pp. 10–14.

[24] J. Berg. “The contrasting and conflicting definitions of envelopment.” In: *Proceedings of the 126th AES Convention, paper no. 7808*. Munich, Germany, 2009.

[25] J. Berg and F. Rumsey. “Identification of quality attributes of spatial audio by repertory grid technique.” In: *Journal of the Audio Engineering Society* 54.5 (2006), pp. 365–379.

[26] A. Berkhout. “A holographic approach to acoustic control.” In: *Journal of the Audio Engineering Society* 36.12 (1988), pp. 977–995.

[27] B. Bernschütz. “A Spherical Far Field HRIR/HRTF Compilation of the Neumann KU 100.” In: *Proceedings of the AIA-DAGA*. Merano, Italy, 2013, pp. 592–595.

[28] B. Bernschütz. “Microphone Arrays and Sound Field Decomposition for Dynamic Binaural Recording.” PhD thesis. Technische Universität Berlin, Germany, 2016.

[29] B. Bernschütz. *Sofia - Soundfield analysis toolbox, Retrieved October 24, 2013.* 2013. URL: https://code.google.com/p/sofia-toolbox/wiki.

[30] B. Bernschütz, C. Pörschmann, S. Spors, and S. Weinzierl. “SOFiA sound field analysis toolbox.” In: *Proceedings of the International Conference on Spatial Audio (ICSA)*. Detmold, Germany, 2011.

[31] S. Bertet, J. Daniel, E. Parizet, and O. Warusfel. “Investigation on the restitution system influence over perceived Higher Order Ambisonics sound field: a subjective evaluation involving from first to fourth order systems.” In: *Journal of the Acoustical Society of America* 123.5 (2008), p. 3936. DOI: 10.1121/1.2936007.

[32] C.M. Bishop. *Neural Networks for Pattern Recognition.* Oxford: Clarendon Press, 1995. ISBN: 978-0198-53864-6.

[33] C.M. Bishop. *Pattern Recognition and Machine Learning.* New York: Springer, 2006. ISBN: 978-0387-31073-2.

[34] J Blauert. “Analysis and synthesis of auditory scenes.” In: *Communications Acoustics.* Ed. by J Blauert. Berlin–Heidelberg–New York: Springer, 2005, pp. 1–26. ISBN: 978-3-540-27437-7.

[35] J. Blauert. *Spatial hearing - The Psychophysics of Human Sound Localization.* Stuttgart, Germany: Hirzel Verlag, 1974.

[36] J. Blauert, J. Braasch, J. Buchholz, H.S. Colburn, U. Jekosch, A.G. Kohlrausch, J. Mourjopoulos, V. Pulkki, and A. Raake. “Aural assessment by means of binaural algorithms -the AABBA project.” In: *Binaural processing and spatial hearing (International Symposium on Auditory and Audiological Research, ISAAR 2009, 26-28 August 2009, Helsingor, Denmark).* Ed. by J. Buchholz, T. Dau, J. Christensen-Dalsgaard, and T. Poulsen. Ballerup, Denmark: The Danavox Foundation, 2010, pp. 113–124.

[37] D.R. Box and G.E.P. Cox. “An Analysis of Transformations.” In: *Journal of the Royal Statistical Society. Series B (Methodological)* 26.2 (1964), pp. 211–252.

[38] J. Bradley, R. Reich, and S. Norcross. “On the combined effects of early- and late-arriving sound on spatial impression in concert halls.” In: *Journal of the Acoustical Society of America* 108.2 (2000), pp. 651–661. DOI: 10.1121/1.429597.

[39] J.S. Bradley and G.A. Soulodre. “The influence of late arriving energy on spatial impression.” In: *Journal of the Acoustical Society of America* 97.4 (1995), pp. 2263–2271. DOI: 10.1121/1.411951.

[40] K. Brandenburg. “Evaluation of quality for audio encoding at low bit rates.” In: *Proceedings of the 82nd AES Convention, paper 2433.* London, UK, 1987.

[41] K. Brandenburg. “OCF - A new coding algorithm for high quality sound signals.” In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).* Vol. 12. 1987, pp. 141–144. DOI: 10.1109/ICASSP.1987.1169893.

[42] K. Brandenburg, F. Klein, A. Neidhardt, and S. Werner. “Auditory Illusion over Headphones Revisited.” In: *Proceedings of the Internation al Conference on Acoustics (ICA).* Boston, USA, 2017.

[43] K. Brandenburg and T. Sporer. “NMR and Masking Flag: Evaluation of quality using perceptual criteria.” In: *Proceedings of the 11th AES Conference, paper 11-020.* Portland, USA, 1992, pp. 169–179.

[44] M. Brandstein and D. Ward. *Microphone Arrays: Signal Processing Techniques and Applications.* 5. Auflage. Frankfurt/Main: Springer, 2010. ISBN: 9783642075476.

[45] L. Breiman. “Bagging predictors.” In: *Machine Learning* 24.2 (1996), pp. 123–140. DOI: 10.1007/BF00058655.

[46] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees.* New York, USA: Chapman and Hall, 1984. ISBN: 978-1351460484.

[47] J. Bridle. “Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition.” In: *Neurocomputing: Algorithms, Architectures and Applications.* Ed. by F. Fogelman Soulie and J. Herault. Berlin: Springer, 1990, pp. 227–236. ISBN: 978-3-642-76153-9.

[48] F. Brinkmann, A. Lindau, M. Vrhovnik, and S. Weinzierl. “Assessing the Authenticity of Individual Dynamic Binaural Synthesis.” In: *Proceedings of the EAA Joint Auralization and Ambisonics Symposium.* Berlin, Germany, 2014, pp. 62–68. DOI: 10.14279/depositonce-11.

[49] F. Brinkmann, R. Roden, A. Lindau, and S. Weinzierl. “Interpolation of head-above-torso-orientations in head-related transfer functions.” In: *Proceedings of the EAA Forum Acousticum.* Krakow, Poland, 2014.

[50] ITU Recommendation ITU-R BS.1116-1. *Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems.* 1997.

[51] ITU Recommendation ITU-R BS.1284-1. *General methods for the subjective assessment of sound quality.* 2003.

[52] ITU Recommendation ITU-R BS.1534-1. *Method for the subjective assessment of intermediate quality level of coding systems.* 2003.

[53] D. Campbell, E. Jones, and M. Glavin. “Audio quality assessment techniques—A review, and recent developments.” In: *Signal Processing* 89.8 (2009), pp. 1489–1500. DOI: 10.1016/j.sigpro.2009.02.015.

[54] B. Cardenas, A. Schmitt, and M. Vance. *A Study of Auditory Source Width and Listener Envelopment.* Tech. rep. New York, USA: School of Architecture at Rensselaer Polytechnic Institute, 2012.

[55] W.J. Cavanaugh and J.A. Wilkes. *Architectural Acoustics: Principles and Practice.* 2nd ed. John Wiley & Sons, 2010. ISBN: 978-0-470-19052-4.

[56] J.M. Chambers, W.S. Cleveland, B. Kleiner, and P.A. Tukey. *Graphical Methods for Data Analysis.* Boston: Duxbury Press, 1983.

[57] S. Choisel and F. Wickelmaier. “Evaluation of multichannel reproduced sound: Scaling auditory attributes underlying listener preference.” In: *Journal of the Acoustical Society of America* 121.1 (2007), pp. 388–400. DOI: 10.1121/1.2385043.

[58] S. Choisel and F. Wickelmaier. “Relating auditory attributes of multichannel sound to preference and to physical parameters.” In: *Proceedings of the 120th AES Convention - paper no. 6684.* Paris, France, May 2006.

[59] W. Cleveland and S. Devlin. "Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting." In: *Journal of the American Statistical Association* 83.403 (1988), pp. 569–610. DOI: 10.1080/01621459.1988.10478639.

[60] R. K. Clifton, R. L. Freyman, R. Y. Litovsky, and D. McCall. "Listeners' expectations about echoes can raise or lower echo threshold." In: *Journal of the Acoustical Society of America* 95.3 (1994), pp. 1525–1533. DOI: 10.1121/1.408540.

[61] J. Cohen. "A Coefficient of Agreement for Nominal Data." In: *Educational and Psychological Measurement* 20.1 (1960), pp. 37–46. DOI: 10.1177/001316446002000104.

[62] C. Colomes, M. Lever, J.-B. Rault, Y.-F. Dehery, and G. Faucon. "A perceptual model applied to audio bit-rate reduction." In: *Journal of the Audio Engineering Society* 43.4 (1995), pp. 233–240.

[63] R. Conetta. "Towards the automatic assessment of spatial quality in the reproduced sound environment." PhD thesis. University of Surrey, 2011.

[64] R. Conetta, T. Brookes, F. Rumsey, S. Zielinski, M. Dewhirst, B. Jackson, S. Bech, D. Mares, and S. George. "Spatial Audio Quality Perception (Part 1): Impact of Commonly Encountered Processes." In: *Journal of the Audio Engineering Society* 62.12 (2014), pp. 831–846. DOI: 10.17743/jaes.2014.0084.

[65] R. Conetta, T. Brookes, F. Rumsey, S. Zielinski, M. Dewhirst, B. Jackson, S. Bech, D. Mares, and S. George. "Spatial Audio Quality Perception (Part 2): A linear Regression Model." In: *Journal of the Audio Engineering Society* 62.12 (2014), pp. 847–860. DOI: 10.17743/jaes.2014.0047.

[66] H. Coolican. *Research Methods and Statistics in Psychology.* 5th ed. London, UK: J.W. Arrowsmith Ltd., 2009. ISBN: 978-0-340-98344-7.

[67] H. Cox, R.M. Zeskind, and M.M. Owen. "Robust adaptive beamforming." In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 35.10 (October 1987), pp. 1365–1376. DOI: 10.1109/TASSP.1987.1165054.

[68] L. Cremer. *Die wissenschaftlichen Grundlagen der Raumakustik (Scientific basics of room acoustics).* 1st ed. Stuttgart, Germany: Hirzel-Verlag, 1948.

[69] L. Cremer and H. Muller. *Principles and Applications of Room Acoustics.* London: Applied Sciences, 1982.

[70] P. Damaske. "Head-related two-channel stereophony with loudspeaker reproduction." In: *Journal of the Acoustical Society of America* 50 (1971), pp. 1109–1115. DOI: 10.1121/1.1912742.

[71] P. Damaske and B. Wagner. “Richtungshörversuche mit nachgebildeten Kopf (Localization experiments based on an artificial head).” In: *Acoustica* 21 (1969), pp. 30–35. DOI: 10.1121/1.1912742.

[72] J. Daniel. “Acoustic field representatio, application to the transmission and the reproduction of complex sound evnironments in a multimedia context.” PhD thesis. University of Paris, 2001.

[73] L.S. Davis, R. Duraiswami, E. Grassi, N.A. Gumerov, Z. Li, and D.N. Zotkin. “High Order Spatial Audio Capture and Its Binaural Head-Tracked Playback Over Headphones with HRTF Cues.” In: *Proceedings of the 119th AES Convention - paper no. 6540*. New York, USA, 2005.

[74] J. Delarue and J.-M. Sieffermann. “Sensory mapping using Flash profile. Comparison with a conventional descriptive method for the evaluation of the flavour of fruit dairy products.” In: *Food Quality and Preference* 15.4 (2004), pp. 383–392. DOI: 10.1016/S0950-3293(03)00085-5.

[75] M. Dietz, S. Ewert, and V. Hohmann. “Auditory model based direction estimation of concurrent speakers from binaural signals.” In: *Speech Communication (53), 592-605* 53.5 (2011), pp. 592–605. DOI: 10.1016/j.specom.2010.05.006.

[76] Christian Doepping. “Untersuchungen zur Auralisation von Arraymessungen unter Berücksichtigung realer Messbedingungen.” In: *Diploma Thesis Fraunhofer IDMT* (2009).

[77] C.L. Dolph. “A current distribution for broadside arrays which optimizes the relationship between beam width and sidelobe level.” In: *Proceedings of the Institute of Radio Engineers (IRE)* 34 (1946), pp. 335–348.

[78] J. van Dorp Schuitmann. “Auditory modelling for assessing room acoustics.” PhD thesis. Technische Universität Delft, 2011.

[79] European Broadcasting Union EBU. *EBU website - Sound Quality Assessment Material recordings for subjective tests*. URL: http://https://tech.ebu.ch/publications/sqamcdl.

[80] V. Erbes, F. Schultz, A. Lindau, and S. Weinzierl. “An extraaural headphone for optimized binaural reproduction.” In: *Proceedings of the 38th DAGA*. Darmstadt, Germany, 2012.

[81] M. Evans, J. Angus, and A. Tew. “Analyzing head-related transfer function measurements using surface spherical harmonics.” In: *Journal of the Acoustical Society of America* 104 (1998), p. 2400. DOI: 10.1121/1.423749.

[82] P. Evjen, J. Bradley, and S. Norcross. “The effect of lateral reflections from above and behind on listener envelopment.” In: *Applied Acoustics* 62 (2001), pp. 137–153.

[83] C.F. Eyring. “Reverberation Time in Dead Rooms.” In: *Journal of the Acoustical Society of America* 1 (1930), pp. 217–241. DOI: 10.1121/1.1901884.

[84] A. Farina. “Advancements in impulse response measurements by sine sweeps.” In: *Proceedings of the 122nd AES Convention, paper no. 7121.* Vienna, Austria, 2007.

[85] A. Farina, P. Martignon, A. Capra, and S. Fontana. “Measuring Impulse Responses Containing Complete Spatial Information.” In: *Proceedings of the 22nd AES Conference - paper no. 9.* Cambridge, UK, 2007.

[86] D. Fitzroy. “Reverberation formulae which seems to be more accurate with non-uniform distributionof absorption.” In: *Journal of the Acoustical Society of America* 31 (1959), pp. 893–897. DOI: 10.1121/1.1907814.

[87] J. Fleßner, S.D. Ewert, B. Kollmeier, and R. Huber. “Application of a binaural perception model for spatial audio quality assessment.” In: *Proceedings of the International Conference on Spatial Audio (ICSA).* Erlangen, Germany, 2014, pp. 26–31.

[88] H. Fletcher. “Auditory patterns.” In: *Reviews of Modern Physics* 12.1 (1940), pp. 47–65. DOI: 10.1103/RevModPhys.12.47.

[89] N. Ford, F. Rumsey, and T. Nind. “Evaluating spatial attributes of reproduced audio events using graphical assessment language – understanding differences in listener depictions.” In: *Proceedings of the 24th AES Conference - paper no. 15.* Alberta, Canada, 2003.

[90] Y. Freund and R.E. Schapire. “A decision-theoretic generalization of on-line learning and an application to boosting.” In: *Journal of Computer and System Sciences* 55.1 (1997), pp. 119–139. DOI: 10.1006/jcss.1997.1504.

[91] J. Friedman. “Multivariate Adaptive Regression Splines.” In: *The Annals of Statistics* 19.1 (1991), pp. 1–141. DOI: 10.1214/aos/1176347963.

[92] J.H. Friedman. *Multivariate adaptive regression splines.* Tech. rep. Department of Statistics, Stanford University, 1990. URL: http://www.slac.stanford.edu/pubs/slacpubs/4750/slac-pub-4960.pdf.

[93] J.H. Friedman. "Regularized Discriminant Analysis." In: *Journal of the American Statistical Association* 84.405 (1989), pp. 165–175. DOI: 10.2307/2289860.

[94] J.H. Friedman. "Stochastic gradient boosting." In: *Computational Statistics & Data Analysis* 38.4 (2002), pp. 367–378. DOI: 10.1016/S0167-9473(01)00065-2.

[95] H. Furuya, K. Fujimoto, C. Young Ji, and N. Higa. "Arrival direction of late sound and listener envelopment." In: *Applied Acoustics* 62 (2001), pp. 125–136.

[96] G.D. Galdo. "Geometry-based Channel Modeling for Multi-User MIMO Systems and Applications." PhD thesis. TU Ilmenau, 2007.

[97] S. Gannot, D. Burshtein, and E. Weinstein. "Signal enhancement using beamforming and nonstationarity with applications to speech." In: *IEEE Transactions Signal Processing* 49.8 (2001), pp. 1614–1626. DOI: 10.1109/78.934132.

[98] C.D. Geisler. *From Sound to Synapse - Physiology of the Mammalian Ear*. Oxford, UK: Oxford University Press, 1998. ISBN: 978-0-19-510025-9.

[99] S. George, S. Zielinski, and F. Rumsey. "Feature Extraction for the Prediction of Multichannel Spatial Audio Fidelity." In: *IEEE Transanctions on Audio, Speech, and Language Processing* 14.6 (2006), pp. 1994–2005. DOI: 10.1109/TASL.2006.883248.

[100] S. George, S. Zielinski, F. Rumsey, and S. Bech. "Evaluating the sensation of envelopment arising from 5-channel surround sound recordings." In: *Proceedings of the 124th AES Convention - paper no. 7382*. Amsterdam, Netherlands, 2008.

[101] M.A. Gerzon. "The Design of Precisely Coincident Microphone Arrays for Stereo and Surround Sound." In: *AES Preprint presented at the 50th AES Convention, London, UK* (1975), pp. L–20.

[102] GFaI. *Akustische Kamera, Retrieved October 24, 2013*. 2013. URL: http://www.gfai.de/deutsch/produkte/signalverarbeitung-akustische-kamera-produkte/akustische-kamera.html.

[103] E.B. Goldstein. *Sensation and Perception*. 7th ed. Belmont, California, USA: Thomson Wadsworth, 2007.

[104] L. Goncalves, A. Subtil, M.R. Oliveira, and P. de Zea Bermudez. "ROC Curve Estimation: An Overview." In: *Statistical Journal* 12.1 (2014), pp. 1–20.

[105] D. Griesinger. “IALF-Binaural Measures of Spatial Impression and Running Reverberance.” In: *Proceedings of the 92nd AES Convention - paper no. 3292.* Vienna, Austria, 1992.

[106] D. Griesinger. “Objective measures of spaciousness and envelopment.” In: *Proceedings of the 16th AES Conference, paper no 16-003.* Rovaniemi, Finland, 1999, pp. 27–41.

[107] D. Griesinger. “Spaciousness and envelopment in musical acoustics.” In: *Proceedings of the 101st AES Convention, paper no. 4401.* Los Angeles, USA, 1996.

[108] D. Griesinger. “Speaker Placement, Externalization, and Envelopment in Home Listening Rooms.” In: *Proceedings of the 105th AES Convention - paper no. 4860.* San Francisco, USA, 1998.

[109] S.R. Gunn. *Support Vector Machines for Classification and Regression.* Tech. rep. Faculty of Electronics and Computer Science, University of Southampton, 1998. URL: `http://users.ecs.soton.ac.uk/srg/publications/pdf/SVM.pdf`.

[110] E. Habets, J. Benesty, I. Cohen, and S. Gannot. “New Insights Into the MVDR Beamformer in Room Acoustics.” In: *IEEE Transactions on Audio, Speech, and Language Processing* 18.1 (2009), pp. 158–170. DOI: `10.1109/TASL.2009.2024731`.

[111] J.L. Hall, V. Madisetti, and D. Williams. *The Digital Signal Processing Handbook.* CRC Press, Boca Raton, FL, USA, 1998.

[112] D. Hammershøi and H. Møller. “Sound transmission to and within the human ear canal.” In: *Journal of the Acoustical Society of America* 100.1 (1996), pp. 408–427.

[113] J. Hanley and B. McNeil. “The Meaning and Use of the Area under a Receiver Operating (ROC) Curvel Characteristic.” In: *Radiology* 143.1 (1982), pp. 29–36.

[114] W.W. Hansen. “A New Principle in Directional Antenna Design.” In: *Proceedings of the Institute of Radio Engineers (IRE)* 26.3 (1938), pp. 333–345. DOI: `10.1109/JRPROC.1938.228128`.

[115] R.H. Hardin and N.J.A. Sloane. “Mclaren’s improved snub cube and other new spherical designs in three dimensions.” In: *Discrete & Computational Geometry* 15.4 (1996), pp. 429–441.

[116] T. Hastie and R. Tibshirani. “Discriminant Analysis by Gaussian Mixtures.” In: *Journal of the Royal Statistical Society, Series B.* 58 (1996), pp. 155–176. DOI: `10.2307/2346171`.

[117] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 2nd ed. Berlin: Springer Verlag, 2009. ISBN: 978-0-387-84858-7.

[118] R. Hellman. "Asymmetry of masking between noise and tone." In: *Perception and Psychophysics* 11.3 (1972), pp. 241–246. DOI: 10.3758/BF03206257.

[119] W. Hess and J. Weishäupl. "Replication of Human Head Movements in 3 Dimensions by a Mechanical Joint." In: *Proceedings of the International Conference on Spatial Audios (ICSA).* Erlangen, Germany, 2014.

[120] J.M. Hirst. "Spatial Impression in Multichannel Surround Sound Systems." PhD thesis. University of Salford, 2006.

[121] J. Ho Kim, C. Ki Seo, H. Joo Yoo, and J. Yong Jeon. "The effect of reflectors on Sound strength (G) and IACC in a fan-shape hall." In: *Proceedings of the International Symposium on Room Acoustics (ISRA).* Melbourne, Australia, 2010, pp. 29–31.

[122] H. Hotelling. "Analysis of a complex of statistical variables into principal components." In: *Journal of Educational Psychology* 24 (1933), pp. 417–441. DOI: 10.1037/h0071325.

[123] E. Hulsebos. "Auralization using Wave Field Synthesis." PhD thesis. Technische Universität Delft, 2004.

[124] J. Hyde. "Sound strength in concert halls I: role of the early sound field with objective and subjective measures." In: *Journal of the Acoustical Society of America* 103.5 (1998), pp. 2748–2748. DOI: 10.1121/1.422800.

[125] F.R. Jack and J.R. Piggott. "Free choice profiling in consumer research." In: *Food Quality and Preference* 3.3 (1991), pp. 129–134. DOI: 10.1016/0950-3293(91)90048-J.

[126] D. Jarrett, E. Habets, M. Thomas, and P. Naylor. *Acoustic signal processing in spherical geometries, Retrieved April 24, 2015.* 2015. URL: http://www.commsp.ee.ic.ac.uk/~sap/projects/spherical-microphone-arrays/.

[127] M. Jeffet, N. Shabtai, and B. Rafaely. "Theory and Perceptual Evaluation of the Binaural Reproduction and Beamforming Tradeoff in the Generalized Spherical Array Beamformer." In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24.4 (2016), pp. 708–718. DOI: 10.1109/TASLP.2016.2522649.

[128] L.A. Jeffress. “A place theory of sound localisation.” In: *Journal of Comparative Physiology and Psychology* 41.1 (1948), pp. 35–39. DOI: 10.1037/h0061495.

[129] U. Jekosch. *Voice and Speech Quality Perception – Assessment and Evaluation*. Berlin, Heidelberg, Germany: Springer, 2005. ISBN: 978-3-540-24095-2.

[130] A.J. Jerri. “The Shannon sampling theorem - Its various extensions and applications: A tutorial review.” In: *Proceedings of the IEEE* 65.11 (1977), pp. 1565–1596. DOI: 10.1109/PROC.1977.10771.

[131] I.T. Jolliffe. *Principal Component Analysis*. 2nd ed. New York, USA: Springer, 2002. ISBN: 978-0-387-22440-4.

[132] M. Karjalainen. “A New Auditory Model for the Evaluation of Sound Quality of Audio Systems.” In: *Proceedings of the International Conference on Audio, Speech, and Signal Processing (ICASSP)*. Vol. 10. Florida, USA, 1985, pp. 608–611. DOI: 10.1109/ICASSP.1985.1168376.

[133] W.d.V. Keet. “The influence of early lateral reflections on spatial impression.” In: *Proceedings of the 6th International Congress on Acoustics (ICA)*. Tokyo, Japan, 1968, E53–E56.

[134] K. Kira and L. Rendell. “The Feature Selection Problem: Traditional Methods and a New Algorithm.” In: *Proceedings of the 10th National Conference on Artificial Intelligence*. San Jose, USA, 1992.

[135] F. Klein and S. Werner. “Auditory Adaptation in Spatial Listening Tasks.” In: *Proceedings of the 138th AES Convention - paper no. 9281*. Warsaw, Poland, 2015.

[136] M. Kleiner, B.-I. Dalenbäck, and P. Svensson. “Auralization - An Overview.” In: *Journal of the Audio Engineering Society* 41.11 (1993), pp. 861–875.

[137] S. Klockgether and S. van de Par. “A Model for the Prediction of Room Acoustical Perception Based on the Just Noticeable Differences of Spatial Perception.” In: *Acta Acustica united with Acustica* 100.5 (2014), pp. 964–971. DOI: 10.3813/AAA.918776.

[138] I. Kononenko. “Estimating Attributes: Analysis and Extensions of Relief.” In: *Machine Learning: ECML–94*. Ed. by F. Bergadano and L. De Raedt. Vol. 784. Springer Berlin / Heidelberg, 1994, pp. 171–182.

[139] A. Koretz and B. Rafaely. “Dolph-Chebyshev beampattern design for spherical arrays.” In: *IEEE Transactions on Signal Processing* 57.6 (2009), pp. 2417–2420. DOI: 10.1109/TSP.2009.2015120.

[140] W. Kuhl. "Spaciousness (spatial impression) as a component of total room impression." In: *Acoustica* 40 (1978), pp. 167–181.

[141] M. Kuhn and K Johnson. *Applied Predictive Modeling.* New York: Springer, 2013. ISBN: 978-1-4614-6848-6.

[142] R. Kürer. "Einfaches Messverfahren zur Bestimmung der Schwerpunktzeit raumakustischer Impulsantworten (A simple measurement procedure to determine the center time of room acoustical impulse responses)." In: *Proceedings of the 7th International Congress on Acoustics (ICA).* Budapest, Hungary, 1971.

[143] H. Kuttruff. *Room Acoustics.* 5th Revised Edition. Routledge Chapman & Hall, 2000.

[144] R. Lacatis, A. Gimenez, A. Barba Sevillano, S. Cerda, J. Romero, and R. Cibrian. "Historical and chronological evolution of the concert hall acoustics parameters." In: *The Journal of the Acoustical Society of America* 123.5 (2008), p. 3198. DOI: 10.1121/1.2933348.

[145] M. Lavandier, S. Jelfs, J.F. Culling, A.J. Watkins, A.P. Raimond, and Makin S.J. "Binaural prediction of speech intelligibility in reverberant rooms with multiple noise sources." In: *Journal of the Acoustical Society of America* 131.1 (2012), pp. 218–231. DOI: 10.1121/1.3662075.

[146] P. Laws. "Zum Problem des Entfernungshörens und der In-Kopf-Lokalisiertheit von Hörereignissen [On the problem of distance headring and in-head localization of auditory events inside the head]." PhD thesis. Technische Hochschule Aachen, 1972.

[147] V. Lebedev. *Fortran Code for Lebedev Grids.* URL: http://www.ccl.net/cca/software/SOURCES/FORTRAN/index.shtml.

[148] V. Lebedev. "Quadratures on a sphere." In: *Computational Mathematics and Mathematical Physics* 16.2 (1975), pp. 293–306.

[149] V. Lebedev. "Spherical quadrature formulas exact to orders 25-29." In: *Translated from Sibirskii Math. Zhurnal* 18 (1977), pp. 99–107.

[150] V. Lebedev. "Values of the nodes and weights of ninth to seventeenth order gauss-markov quadrature formulae invariant under the octahedron group with inversion." In: *Computational Mathematics and Mathematical Physics* 15 (1975), pp. 44–51.

[151] H. Lee. “Apparent Source Width and Listener Envelopment in Relation to Source-Listener Distance.” In: *Proceedings of the 52nd International AES Conference, paper 3-1.* Guildford, UK, 2013.

[152] P. Lehmann and H. Wilkens. “Zusammenhang subjektiver Beurteilung von Konzertsälen mit raumakustischen Kriterien (Relation between subjective Assessment of concert halls and room acoustical criteria).” In: *Acoustica* 45.4 (1980), pp. 256–268.

[153] T. Letowski. “Sound Quality Assessment: Concepts and Criteria.” In: *Proceedings of the 87th AES Convention, preprint no. 2825.* New York, USA, 1989.

[154] Z. Li and R. Duraiswami. “Flexible and optimal desing of spherical microphone arrays for beamforming.” In: *IEEE Transactions on Audio, Speech, and Language Processing* 15.2 (2007), pp. 702–714. DOI: 10.1109/TASL.2006.876764.

[155] J. Liebetrau, S. Kämpf, S. Schneider, and T. Sporer. “Standardization of PEAQ-MC: Extension of ITU-R BS.1387-1 to Multichannel Audio.” In: *Proceedings of the 40th AES Conference, paper no. P-3.* Tokyo, Japan, 2010.

[156] A. Lindau. “Binaural Resynthesis of Acoustical Environments - Technology and Perceptual Evaluation.” PhD thesis. TU Berlin, 2014.

[157] A. Lindau, V. Erbes, S. Lepa, H.-J. Maempel, F. Brinkmann, and S. Weinzierl. “A Spatial Audio Quality Inventory (SAQI).” In: *Acta Acoustica united with Acoustica* 100 (2014), pp. 984–994. DOI: 10.3813/AAA.918778.

[158] A. Lindau and S. Weinzierl. “FABIAN - An instrument for software-based measurement of binaural room impulse responses in multiple degrees of freedom.” In: *Proceedings of the 24th Tonmeistertagung.* Leipzig, Germany, 2006, pp. 621–625.

[159] T. Lokki and H. Järveläinen. “Subjective evaluation of auralization of physics-based room acoustics modeling.” In: *Proceedings of the 7th International Conference on Auditory Display (ICAD).* Espoo, Finnland, 2001.

[160] T. Lokki, J. Pätynen, A. Kusinen, and S. Tervo. “Disentangling preference ratings of concert hall acoustics using subjective sensory profiles.” In: *Journal of the Acoustic Society of America* 132 (2012), pp. 3148–3161. DOI: 10.1121/1.4756826.

[161] M. Lombard and T. Ditton. “At the Heart of It All: The Concept of Presence.” In: *Journal of Computer Mediated-Communication* 3.2 (1997). DOI: 10.1111/j.1083-6101.1997.tb00072.x.

[162] G. Lorho. “Individual Vocabulary Profiling of Spatial Enhancement Systems for Stereo Headphone Reproduction.” In: *Proceedings of the 119th AES Convention - paper preprint no. 6629*. New York, USA, 2005.

[163] P. Mackensen. “Auditive Localization. Head Movements, an additional cue in Localization.” PhD thesis. TU Berlin, 2004.

[164] P. Majdak, P. Søndergaard, and J. Blauert. “Die Auditory-Modeling Toolbox der AABBA-Initiative.” In: *Proceedings of the DAGA*. Darmstadt, Germany, 2014, pp. 584–585.

[165] D.G. Malham. “Higher order Ambisonic systems for the spatialisation of sound.” In: *Proceedings of the International Computer Music Conference (ICMC)*. Beijing, China, 1999, pp. 484–487.

[166] A. Marshall and M. Barron. “Spatial responsiveness in concert halls and the origins of spatial impression.” In: *Applied Acoustics* 62 (2001), pp. 91–108.

[167] F. Melchior. “Investigations on spatial sound design based on measured room impulse responses.” PhD thesis. TU Delft, 2011.

[168] F. Melchior, Z. Kuang, D de Vries, and S. Brix. “Spherical Array Systems - On the effect of measurement errors in terms of perceived auralization quality.” In: *Proceedings of the NAG/DAGA*. Rotterdam, Denmark, 2009.

[169] F. Melchior, O. Thiergart, D. de Vries, and S. Brix. “Dual radius spherical cardioid microphone arrays for binaural auralization.” In: *Proceedings of the 127th AES Convention, paper no. 7855*. New York, USA, 2009.

[170] J. Meyer and G. Elko. “A highly scalable spherical microphone array based on an orthonormal decomposition of the sound field.” In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Vol. 2. Orlando,USA, 2002, pp. II-1781–II-1784. DOI: 10.1109/ICASSP.2002.5744968.

[171] J. Meyer and G. Elko. “A spherical microphone array for spatial sound recordings.” In: *Journal of the Acoustical Society of America* 111.5/2 (2002), pp. 2346–2346. DOI: 10.1121/1.4809138.

[172] J.C. Middlebrooks, J.C. Makous, and D.M. Green. “Directional sensitivity of the sound-pressure level in the human ear canal.” In: *Journal of the Acoustical Society of America* 86.1 (1989), pp. 89–108.

[173] T.W. Miller. *Modeling Techniques in Predictive Analytics with Python and R - A Guide to Data Science.* Upper Saddle River, New Jersey, USA: Pearson Education, Inc., 2015. ISBN: 978-0-13-389206-2.

[174] G. Millingtone. “A Modified Formula for Reverberation.” In: *Journal of the Acoustical Society of America* 4 (1932), pp. 69–82. DOI: 10.1121/1.1915588.

[175] P. Minnaar, S.K. Olesen, F. Christensen, and H. Møller. “Localization with Binaural Recordings from Artificial and Human Heads.” In: *Journal of the Audio Engineering Society* 49.5 (2001), pp. 323–336.

[176] C. Moldrzyk. “Ein neuartiger Kunstkopf zur Verifikation einer akustischen Entwurfsmethodik für Architekten (A novel dummy head for verification of an acoustical development method for architects).” In: *Proceedings of the 22nd Tonmeistertagung.* Hannover, Germany, 2002.

[177] H. Møller. “Fundamentals of Binaural Technology.” In: *Applied Acoustics* 36 (1992), pp. 171–218. DOI: 10.1016/0003-682X(92)90046-U.

[178] H. Møller, C.B. Jensen, D. Hammershøi, and M.-F. Sørensen. “Using a Typical Human Subject for Binaural Recording.” In: *Proceedings of the 100th AES Convention - paper no. 4157.* Copenhagen, Denmark, 1996.

[179] H. Møller, M.-F. Sørensen, D. Hammershøi, and C.B. Jensen. “Head-Related Transfer Functions of Human Subjects.” In: *Journal of the Audio Engineering Society* 43.5 (1995), pp. 300–321.

[180] H. Møller, M.-F. Sørensen, C.B. Jensen, and D. Hammershøi. “Binaural Technique: Do We Need Individual Recordings?” In: *Journal of the Audio Engineering Society* 44.6 (1996), pp. 451–469.

[181] B. Moore. *An Introduction to the Psychology of Hearing.* 5th ed. Academic Press, 2003. ISBN: 978-0125056281.

[182] B.C.J. Moore and Glasberg B. “Suggested formulae for calculating auditory-filter bandwidths and excitation patterns.” In: *Journal of the Acoustical Society of America* 74.3 (1983), pp. 750–753. DOI: 10.1121/1.389861.

[183] M. Morimoto. “The relation between spatial impression and the precedence effect.” In: *Proceedings of the 8th International Conference on Auditory Display.* Kyoto, Japan, 2002.

[184] M. Morimoto, K. Lida, and K. Sakagami. “The role of reflections from behind the listener in spatial impression.” In: *Applied Acoustics* 62 (2001), pp. 91–108.

[185] M. Morimoto and Z. Maekawa. "Auditory spaciousness and envelopment." In: *Proceedings of the 13th International Conference on Acoustics (ICA)*. Belgrade, Serbia, 1989.

[186] M. Möser. *Technische Akustik*. 6th ed. Berlin Heidelberg: Springer Verlag, 2005. ISBN: 978-3-540-26445-3.

[187] G. Müller and M. Möser. *Taschenbuch der technischen Akustik*. 3rd ed. Berlin: Springer Verlag, 2004. ISBN: 978-3540412427.

[188] U. Neisser. *Cognition and Reality: Principles and Implications of Cognitive Psychology*. San Francisco, California, USA: W.H. Freeman, 1976.

[189] J. Nelder and R. Wedderbum. "Generalized Linear Models." In: *Journal of the Royal Statistical Society* 135.3 (1972), pp. 370–384. DOI: 10.2307/2344614.

[190] R. Nicol. *Binaural Technology*. Audio Engineering Society Inc, New York, USA, 1999.

[191] R. Nicol, L. Gros, C. Colomes, M. Noistering, O. Warusfel, H. Bahu, B. Katz, and L. Simon. "A Roadmap for Assessing the Quality of Experience of 3D Audio Binaural Rendering." In: *Proceedings of the EAA Joint Symposium on Auralization and Ambisonics - paper no: 16*. Berlin, Germany, Mar. 2014. DOI: 10.14279/depositonce-17.

[192] J. Nowak. "Modeling the Perception of System Errors in Spherical Microphone Array Auralizations." In: *Journal of the Audio Engineering Society (submitted)* (2019).

[193] J. Nowak, K. Jurgeit, and J. Liebetrau. "Assessment of spherical microphone array auralizations using open-profiling of quality (OPQ)." In: *Proceedings of the 8th International Workshop on Quality of Multimedia Experience (QoMEX)*. Lisbon, Portugal, 2016, pp. 1–6. DOI: 10.1109/QoMEX.2016.7498928.

[194] J. Nowak and S. Klockgether. "Perception and prediction of apparent source width and listener envelopment in binaural spherical microphone array auralizations." In: *Journal of the Acoustical Society of America* 142.3 (2017), pp. 1634–1645. DOI: 10.1121/1.5003917.

[195] J. Nowak, J. Liebetrau, and T. Sporer. "On the perception of apparent source width and listener envelopment in wave field synthesis." In: *Proceedings of the 5th International Workshop on Quality of Multimedia Experience (QoMEX)*. Klagenfurth, Austria, 2013, pp. 82–87. DOI: 10.1109/QoMEX.2013.6603215.

[196] D. Nyberg and J. Berg. “Listener Envelopment - What has been done and what future research is needed?” In: *Proceedings of 124th AES Convention, preprint no. 7379*. Amsterdam, Netherlands, 2008.

[197] T. Okano, L. Beranek, and T. Hidaka. “Relations among interaural cross-correlation coefficient ($IACC_E$), lateral fraction (LF_E), and apparent source width (ASW) in concert halls.” In: *Journal of the Acoustical Society of America* 104.1 (1998), pp. 255–265. DOI: 10.1121/1.423955.

[198] H. Okubo, M. Otani, R. Ikezawa, and K. Komiyama S. Nakabayashi. “A system for measuring the directional room acoustical parameters.” In: *Applied Acoustics* 62 (2001), pp. 203–215.

[199] ITU Recommendation ITU-T P.800. *Methods for subjective determination of transmission quality*. 1996.

[200] ITU Recommendation ITU-T P.862. *Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*. 2001.

[201] ITU Recommendation ITU-T P.862.2. *Wideband extension to recommendation for the assessment of wideband telephone networks and speech codecs*. 2005.

[202] B. Paillard, P. Mabilleau, S. Morissette, and J. Soumagne. “PERCEVAL: perceptual evaluation of the quality of audio signals.” In: *Journal of the Audio Engineering Society* 40.1/2 (1992), pp. 21–31.

[203] M. Paquier, V. Koehl, R. Nicol, and J. Daniel. “Subjective assessment of microphone arrays for spatial audio recording.” In: *Proceedings of the Forum Acousticum*. Aalborg, Denmark, 2011.

[204] R.S. Pellegrini. “A virtual reference listening room as an application of auditory virtual environments.” PhD thesis. Ruhr University Bochum, 2001.

[205] J.C.Y. Peng. *The Fourier Integral and Its Applications*. London, UK: SAGE Publications, Inc., 1962. ISBN: 978-1-412-95674-1.

[206] M.S. Pepe, G. Longton, and H. Janes. “Estimation and Comparison of Receiver Operating Characteristic Curves.” In: *The Stata Journal* 9.1 (2009), pp. 1–16.

[207] G. Plenge. “On the differences between localization and lateralization.” In: *Journal of the Acoustical Society of America* 56 (1974), pp. 944–951. DOI: 10.1121/1.1903353.

[208] V. Pulkki. “Evaluating spatial sound with binaural auditory model.” In: *Proceedings of the International Computer Music Conference (ICAD).* Havanna, Cuba, June 2007, pp. 73–76.

[209] V. Pulkki, M. Karjalainen, and J. Huopaniemi. “Analyzing Virtual Sound Source Attributes Using a Binaural Auditory Model.” In: *Journal of the Audio Engineering Society* 47.4 (1999), pp. 203–2017.

[210] Nicol. R., H. Shaiek, and P. Rueff. “Objective and subjective assessment of various headphones for spatial audio rendering.” In: *Proceedings of the Forum Acousticum.* Aalborg, Denmark, 2011.

[211] A. Raake. *Speech Quality of VoIP: Assessment and Prediction.* Wiley & Sons, Chichester, UK: Springer, 2006. ISBN: 978-0-470-03060-8.

[212] A. Raake and J. Blauert. “Comprehensive modeling of the formation process of sound quality.” In: *Proceedings of the 5th International Workshop on Quality of Multimedia Experience (QoMEX).* Klagenfurt, Austria, 2013, pp. 76–81.

[213] A. Raake and B. Katz. “Measurement and Prediction of Speech Intelligibility in a Virtual Chat Room.” In: *Proceedings of the 2nd ISCA/DEGA Tutorial and Research Workshop on Perceptual Quality of Systems.* Berlin, Germany, 2006, pp. 40–43.

[214] A. Raake, H. Wierstorf, and J. Blauert. “A case for TWO!EARS in audio quality assessment.” In: *Proceedings of the Forum Acousticum.* Krakow, Poland, 2014.

[215] B. Rafaely. “Analysis and design of spherical microphone arrays.” In: *IEEE Transaction on Speech and Audio Processing* 13.1 (2005), pp. 134–143. DOI: 10.1109/TSA.2004.839244.

[216] B. Rafaely. “Bessel Nulls Recovery in Spherical Microphone Arrays for Time-Limited Signals.” In: *IEEE Transactions on Audio Speech and Language Processing* 19.8 (2011), pp. 2430–2438. DOI: 10.1109/TASL.2011.2136338.

[217] B. Rafaely. “Decomposition of reverberant sound fields into plane waves using microphone arrays.” In: *Proceedings of the 5th International Workshop on Microphone Array Systems.* Erlangen-Nuremberg, Germany, 2003.

[218] B. Rafaely. *Fundamentals of Spherical Array Processing.* Heidelberg, Germany: Springer, 2015. ISBN: 978-3-662-45663-7.

[219] B. Rafaely. "Phase-Mode versus Delay-and-Sum Spherical Microphone Array Processing." In: *IEEE Signal Processing Letters* 12.10 (2005), pp. 713–716. DOI: 10.1109/LSP.2005.855542.

[220] B. Rafaely. "Plane-wave decomposition of the pressure on a sphere by spherical convolution." In: *Journal of the Acoustical Society of America* 116.4 (2004), pp. 2149–2157. DOI: 10.1121/1.1792643.

[221] B. Rafaely and A. Avni. "Interaural cross correlation in a sound field represented by spherical harmonics." In: *Journal of the Acoustical Society of America* 127.2 (2010), pp. 823–828. DOI: 10.1121/1.3278605.

[222] B. Rafaely, Weissm B., and E. Bachmat. "Spatial aliasing in spherical microphone arrays." In: *IEEE Transactions on Signal Processing* 55.3 (2007), pp. 1003–1010. DOI: 10.1109/TSP.2006.888896.

[223] B. Rafaely, I. Balmages, and L. Eger. "High-resolution plane-wave decomposition in an auditorium using a dual-radius scanning spherical microphone array." In: *Journal of the Acoustical Society of America* 122.5 (2007), pp. 2661–2668. DOI: 10.1121/1.2783204.

[224] B. Rafaely and M. Park. "Plane-wave decomposition by spherical-convolution microphone array." In: *Journal of the Acoustical Society of America* 115.5.2 (2004), pp. 2578–2578. DOI: 10.1121/1.4784249.

[225] B. Rafaely and M. Park. "Super-resolution spherical microphone arrays." In: *Proceedings of the 23rd IEEE Convention of Electrical and Electronics Engineers in Israel.* Herzelia, Israel, 2004, pp. 424–427. DOI: 10.1109/EEEI.2004.1361182.

[226] International Telecommunication Union Recommendation. "Method for objective measurements of perceived audio quality." In: *ITU-R BS.1387-1.* Geneva, Switzerland, 2001.

[227] D. Reshef, Y. Reshef, H. Finucane, S. Grossman, G. McVean, P. Turnbaugh, E. Lander, M. Mitzenmacher, and P. Sabeti. "Detecting Novel Associations in Large Data Sets." In: *Science* 334.6062 (2011), pp. 1518–1524. DOI: 10.1126/science.1205438.

[228] B. Ripley, B. Venables, B.M. Bates, K. Hornik, A. Gebhardt, and D. Firth. *Support Functions and Datasets for Venables and Ripley's MASS*. Tech. rep. CRAN r-project, 2015. URL: https://cran.r-project.org/web/packages/MASS/MASS.pdf.

[229] B.D. Ripley. *Pattern Recognition and Neural Networks.* Cambridge, UK: Cambridge University Press, 1996. ISBN: 978-0521717700.

[230] A.W. Rix, J.G. Beerends, D.-S. Kim, P. Kroon, and O. Ghitza. "Objective Assessment of Speech and Audio Quality – Technology and Applications." In: *IEEE Transactions on audio, speech, and language processing* 14.6 (2006), pp. 1890–1901. DOI: 10.1109/TASL.2006.883260.

[231] D. Robinson. "Perceptual model for assessment of coded audio." PhD thesis. University of Essex, 2002.

[232] A. Roginska and P. Geluso. *Immersive Sound: The Art and Science of Binaural and Multi-Channel Audio.* Waltham, Massachusetts, USA: Focal Press, 2017. ISBN: 9781138900004.

[233] T. Rohdenburg, S. Goetze, V. Hohmann, K.-D. Kammeyer, and B. Kollmeier. "Objective perceptual quality assessment for self-steering binaural hearing aid microphone arrays." In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* Las Vegas, USA, 2008, pp. 2449–2452.

[234] F. Rumsey. *Spatial audio.* Oxford, England: Focal Press, 2001.

[235] F. Rumsey. "Spatial Quality Evaluation for Reproduced Sound: Terminology, Meaning, and a Scene-Based Paradigm." In: *Journal of the Audio Engineering Society* 50.9 (2002), pp. 651–666.

[236] F. Rumsey. "Subjective Assessment of the Spatial Attributes of Reproduced Sound." In: *Proceedings of the 15th AES Conference, paper no. 15-012.* Copenhagen, Denmark, 1998.

[237] F. Rumsey, S. Zielinski, R. Kassier, and S. Bech. "On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio quality." In: *Journal of the Acoustical Society of America* 118.2 (2005), pp. 968–976. DOI: 10.1121/1.1945368.

[238] W.C. Sabine. *Collected Papers on Acoustics.* Cambridge: Harvard University Press, 1922.

[239] A. Sarroff. "Subjective Evaluation of Spatial Impression in Reproduced Stereophonic Music." In: *Proceedings of the 6th sound and music computing conference (smc-09).* Porto, Portugal, 2009.

[240] A. Sarroff and J. Bello. “Toward a Computational Model of Perceived Spaciousness in Recorded Music.” In: *Journal of the Audio Engineering Society* 59.7/8 (2011), pp. 498–513.

[241] L. Savioja, J. Huopaniemi, T. Lokki, and R. Väänänen. “Creating interactive virtual acoustic environments.” In: *Journal of the Audio Engineering Society* 47.9 (1999), pp. 675–705.

[242] B. Sayers and E.C. Cherry. “Mechanism of binaural fusion in the hearing of speech.” In: *Journal of the Acoustical Society of America* 29 (1957), pp. 973–987. DOI: 10.1121/1.1914990.

[243] Z. Schärer and A. Lindau. “Evaluation of equalization methods for binaural signals.” In: *Proceedings of the 126th AES Convention - paper no. 7721.* Munich, Germany, May 2009.

[244] B. Scharf. *Foundations of Modern Auditory Theory.* Academic Press, 1970.

[245] M.R. Schröder. “New method of measuring reverberation time.” In: *Journal of the Acoustical Society of America* 37 (1965), pp. 409–409. DOI: 10.1121/1.1909343.

[246] M.R. Schröder, B.S. Atal, and J.L. Hall. “Optimizing digital speech coding by exploiting masking properties of the human ear.” In: *Journal of the Acoustical Society of America* 66 (1979), pp. 1647–1652.

[247] P. Schubert. “Untersuchungen über die Wahmehmbarkeit von Einzelrückwürfen bei Musik (Investigations on the perception of single reflexions in music).” In: *Technische Mitteilungen RFZ* 3 (1966), p. 124.

[248] T. Schubert, F. Friedmann, and H. Regenbrecht. “The Experience of Presence: Factor Analytic Insights.” In: *MIT Press Journals - Presence: Teleoperators and Virtual Environments* 10.3 (2001), pp. 266–281. DOI: 10.1162/105474601300343603.

[249] F. Schultz, A. Lindau, M. Makarski, and S. Weinzierl. “An extraaural headphone for optimized binaural reproduction.” In: *Proceedings of the 26th Tonmeistertagung - VDT International Convention.* Leipzig, Germany, 2010.

[250] T. Schultz. “Acoustics of the Concert Hall.” In: *IEEE Spectrum* 2.6 (1965), pp. 56–67. DOI: 10.1109/MSPEC.1965.5531663.

[251] B. Seeber. “Untersuchung der auditiven Lokalisation mit einer Lichtzeigermethode.” PhD thesis. TU München, 2003.

[252] J.-H. Seo, S.B. Chon, K.-M. Sung, and I. Choi. "Perceptual Objective Quality Evaluation Method for High Quality Multichannel Audio Codecs." In: *Journal of the Audio Engineering Society* 61.7/8 (2013), pp. 535–545.

[253] N. Shabtai and B. Rafaely. "Generalized Spherical Array Beamforming for Binaural Speech Reproduction." In: *IEEE Transactions on Audio, Speech, and Language Processing* 22.1 (2014), pp. 238–247. DOI: 10.1109/TASLP.2013.2290499.

[254] J. Sheaffer, S. Villeval, and B. Rafaely. "Rendering binaural room impulse responses from spherical microphone array recordings using timbre correction." In: *Proceedings of the EAA Joint Symposium on Auralization and Ambisonics*. Berlin, Germany, 2014.

[255] A. Silzle. "3D Audio Quality Evaluation: Theory and Practice." In: *Proceedings of the International Conference on Spatial Audio (ICSA)*. Erlangen, Germany, May 2014.

[256] A. Silzle. "Generation of Quality Taxonomies for Auditory Virtual Environments by Means of Systematic Expert Survey." PhD thesis. Ruhr University Bochum, 2007.

[257] A. Smimite, A. Beghadi, K. Chen, and O. Jafjaf. "A New Approach for Spatial Audio Quality Assessment." In: *Proceedings of the International Conference on Telecommunications and Multimedia (TEMU2014)*. Crete, Greece, July 2014.

[258] S.W. Smith. *Scientist and Engineer's Guide to Digital Signal Processing*. Bertrams, 1997. ISBN: 978-0966017-632.

[259] P. Søndergaard and P. Majdak. "The Auditory Modeling Toolbox." In: *The Technology of Binaural Listening*. Ed. by J. Blauert. Heidelberg–New York–Dordrecht–London: ASA-Press & Springer, 2013, pp. 33–56.

[260] G. Soulodre. "Can Reproduced Sound be Evaluated using Measures Designed for Concert Halls?" In: *Workshop on Spatial Audio and Sensory Evaluation Techniques (SASET)*. Guildford, UK, 2006.

[261] G. Soulodre and J. Bradley. "A subjective evaluation of new acoustic measures in concert halls." In: *Journal of the Acoustical Society of America* 98.1 (1995), pp. 294–301. DOI: 10.1121/1.413735.

[262] J. Spille. "Messung der Vor- und Nachverdeckung bei Impulsen unter kritischen Bedingungen (Measurement of pre- and postmasking of impulses under critical conditions)." In: *Technical report, Thomson Consumer Electronics,* Research and Development Laboratories, Hannover, Germany (unpublished) (1992).

[263] T. Sporer. "Objective audio signal evaluation - applied psychoacoustics for modeling the perceived quality of audio signals." In: *Proceedings of the 103rd AES Convention, preprint 4512.* New York, USA, 1997.

[264] T. Sporer. "Qualitätsbeurteilung von Audiosignalen mittels gehörangepasster Messverfahren." PhD thesis. FAU Erlangen, 1998.

[265] T. Sporer. "Qualitätsbeurteilung von Audiosignalen – Vom Hörtest zum Messverfahren." In: *Proceedings of the 41st DAGA.* Nuremberg, Germany, 2015.

[266] S. Spors, H. Wierstorf, A. Raake, F. Melchior, and F. Zotter. "Spatial Sound With Loudspeakers and Its Perception: A Review of the Current State." In: *Proceedings of the IEEE* 101.9 (2013), pp. 1920–1938. DOI: `10.1109/JPROC.2013.2264784`.

[267] P. Sprent and N.C. Smeeton. *All of Nonparametric Statistics.* 3rd Edition. New York, USA: Chapman & Hall/CRC, 2001. ISBN: 1-58488-145-3.

[268] R.M. Stern, J.G. Brown, and D. Wang. "Binaural Sound Localization." In: *Computational Auditory Scene Analysis. Principals, Algorithms, and Applications.* Ed. by D. Wang and J.G. Brown. Hoboken, NJ: John Wiley & Sons, 2006, pp. 147–185.

[269] D. Strohmeier. "Open Profiling of Quality: A Mixed Methods Research Approach for Audiovisual Quality Evaluations." PhD thesis. TU Ilmenau, 2011.

[270] D. Strohmeier, S. Jumisko-Pyykkö, and K. Kunze. "Open profiling of quality: A mixed method approach to understanding multimodal quality perception." In: *Advances in Multimedia* 2010, Article ID: 658980 (2010), 28 pages. DOI: `10.1155/2010/658980`.

[271] J.W. Strutt. "On Our Perception of Sound Direction." In: *Philosophical Magazine* 6 (1907), pp. 214–232.

[272] J. Stuart. "Noise: methods for estimating detectability and threshold." In: *Proceedings of the 94th AES Convention, paper no. 3477.* Berlin, Germany, 1993.

[273] M. Takanen, H. Wierstorf, V. Pulkki, and A. Raake. "Evaluation of sound field synthesis techniques with a binaural auditory model." In: *Proceedings of the 55th International AES Conference on Spatial Audio - paper no: 6-3.* Helsinki, Finnland, Aug. 2014.

[274] E. Terhardt. “Calculating virtual pitch.” In: *Hearing Research* 1.2 (1979), pp. 155–182. DOI: 10.1016/0378-5955(79)90025-X.

[275] T. Thiede and E. Kabot. “A new perceptual quality measure for bit rate reduced audio.” In: *Proceedings of the 100th AES Convention, preprint 4280.* Copenhagen, Denmark, 1996.

[276] T. Thiede, C.T. William, R. Bitto, C. Schmidmer, T. Sporer, J.G. Beerends, C. Colomes, M. Keyhl, G. Stoll, K. Brandenburg, and B. Feiten. “PEAQ - The ITU standard for objective measurement of perceived audio quality.” In: *Journal of the Audio Engineering Society* 48.1/2 (2000), pp. 3–29.

[277] M. Titterington. “Neural Networks.” In: *Wiley Interdisciplinary Reviews: Computational Statistics* 2.1 (2010), pp. 1–8. DOI: 10.1002/wics.50.

[278] F.E. Toole. “In-Head Localization of Acoustic Images.” In: *Journal of the Acoustical Society of America* 48.4 (1970), pp. 943–949. DOI: 10.1121/1.1912233.

[279] H.L. Van Trees. *Detection, Estimation, and Modulation Theory, Optimum Array Processing.* Wiley, 2004. ISBN: 9780471463832.

[280] V. Vapnik. *The Nature of Statistical Learning Theory.* 2nd ed. Springer: New York, 2000. ISBN: 978-1-4757-3264-1.

[281] V. Vapnik, S. Golowich, and A. Smola. “Support vector method for function approximation, regression estimation, and signal processing.” In: *Advances in Neural Information Processing Systems 9.* Ed. by M. Mozer, M. Jordan, and T. Petsche. MIT Press, Cambridge, USA, 1997, pp. 281–287.

[282] W.N. Venables and B.D. Ripley. *Modern Applied Statistics with S.* 4th ed. Berlin-Heidelberg: Springer, 2002. ISBN: 978-0387954578.

[283] J. Victoria and T. Gorne. “Apparent Source Width in ITU Surround.” In: *Proceedings of the 126th AES Convention, paper no. 7809.* Munich, Germany, 2009.

[284] F Völk. “Psychoacoustic Experiments with Loudspeaker-Based Virtual Acoustics.” In: *Zotter, F. (ed) Chapter 5, Hot Topics in Acoustics: Cutting Edge in Spatial Audio, EAA Documenta Acustica, Merano, Italy.* 2013, pp. 39–44.

[285] F. Völk, F. Heinemann, and H. Fastl. “Externalization in binaural synthesis: effects of recording environment and measurement procedure.” In: *Proceedings of the Acoustics 08.* Paris, France, 2008, pp. 6419–6424.

[286] M. Vorländer. *Auralization - Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality*. Berlin Heidelberg, Germany: Springer-Verlag, 2008. ISBN: 978-3-540-48830-9.

[287] A. Wabnitz, N. Epain, C. Jin, and A. van Schaik. "Room acoustics simulation for multichannel microphone arrays." In: *Proceedings of the International Symposium on Room Acoustics (ISRA)*. Melbourne, Australia, Aug. 2010.

[288] S. Wang, A. Sekey, and A. Gersho. "An objective measure for predicting subjective quality of speech coders." In: *IEEE Journal on Selected Areas in Communications* 10.5 (1992), pp. 819–829. DOI: 10.1109/49.138987.

[289] S. Wang, A. Sekey, and A. Gersho. "Objective estimation of perceived speech quality. I. Development of the measuring normalizing block technique." In: *IEEE Transactions on Speech and Audio Processing* 7.4 (1999), pp. 371–382. DOI: 10.1109/89.771259.

[290] L. Wassermann. *All of Nonparametric Statistics*. Berlin Heidelberg, Germany: Springer-Verlag, 2006. ISBN: 978-0387-25145-5.

[291] S. Weinzierl and H.-J. Maempel. "Sind Hörversuche subjektiv? Zur Objektivität akustischer Maße (Are listening experiments subjective? On the objectivity of acoustical measures)." In: *Proceedings of the DAGA*. Darmstadt, Germany, 2012, pp. 315–316.

[292] S. Weinzierl and M. Vorländer. "Room Acoustical Parameters as Predictors of Room Acoustical Impression: What Do We Know and What Would We Like to Know?" In: *Acoustics Australia* 43.1 (2015), pp. 41–48. DOI: 10.1007/s40857-015-0007-6.

[293] E. M. Wenzel. "What Perception Implies About Implementation of Interactive Virtual Acoustic Environments." In: *Proceedings of the 101st AES Convention - paper preprint no. 4353*. Los Angeles, USA, 1996.

[294] S. Werner and F. Klein. "Influence of Context Dependent Quality Parameters on the Perception of Externalization and Direction of an Auditory Event." In: *Proceedings of the 55th AES Conference - paper no. 6-4*. Helsinki, Finland, 2014.

[295] S. Werner, F. Klein, and K. Brandenburg. "Influence of Scene Complexity and Room Acoustic Disparity on Perception of Quality Features using a Binaural Synthesis System." In: *Proceedings of the 7th International Workshop on Quality*

of Multimedia Experience (QoMEX). Costa Navarino, Messinia, Greece, 2015. DOI: `10.1109/QoMEX.2015.7148131`.

[296] S. Werner and J. Liebetrau. "Adjustment of Direct-to-Reverberant-Energy-Ratio and the Just-Noticable-Difference." In: *Proceedings of the 6th International Workshop on Quality of Multimedia Experience (QoMEX).* Singapore, Singapore, 2014.

[297] H. Wierstorf. "Perceptual Assessment of Sound Field Synthesis." PhD thesis. TU Berlin, 2014.

[298] A.A. Williams and S.P. Langron. "The use of free-choice profiling for the evaluation of commercial ports." In: *Journal of the Science of Food and Agriculture* 35.5 (1984), pp. 558–568. DOI: `10.1002/jsfa.2740350513`.

[299] E. Williams. *Fourier Acoustics - Sound Radiation and Nearfield Acoustical Holography.* San Diego: Academic Press, 1999. ISBN: 9780127539607.

[300] H. Wold. "Estimation of Principal Components and Related Models by Iterative Least Squares." In: *Multivariate Analyses.* Ed. by Krishnaiah P. New York, USA: Academic Press, 1966, pp. 391–420.

[301] D.H. Wolpert. "The Lack of A Priori Distinctions Between Learning Algorithms." In: *Neural Computation* 8.7 (1996), pp. 1341–1390. DOI: `10.1162/neco.1996.8.7.1341`.

[302] D.H. Wolpert and W.G. Macready. "No Free Lunch Theorems for Optimization." In: *IEEE Transactions on Evolutionary Computation* 1.67 (1997), pp. 67–82. DOI: `10.1109/4235.585893`.

[303] H. Yoo Joo, C. Seo Ki, J. Ho Kim, and J. Yong Jeon. "Acoustical Renovation of Large Auditorium to Enhance Sound Strength and IACC." In: *Proceedings of the International Symposium on Room Acoustics (ISRA).* Melbourne, Australia, 2010.

[304] W.A. Yost. *Fundamentals of Hearing - An Introduction.* 4th ed. San Diego, USA: Academic Press, 2000. ISBN: 0-12-775695-7.

[305] J. You, U. Reiter, M.M. Hannuksela, M. Gabbouj, and A. Perkins. "Perceptual-based quality assessment for audio-visual services: A survey." In: *Signal Processing: Image Communication, Special Issue on Image and Video Quality Assessment* 25.7 (2010), pp. 482–501. DOI: `doi:10.1016/j.image.2010.02.002`.

[306] N. Zacharov and K. Koivuniemi. "Audio descriptive analysis and mapping of spatial sound displays." In: *Proceedings of the 7th International Conference on Auditory Displays (ICAD).* Espoo, Finland, 2001, pp. 95–104.

[307] H. Ziegelwanger and P. Majdak. “Modeling the direction-continuous time-of-arrival in head-related transfer functions.” In: *Journal of the Acoustical Society of America* 135.3 (2014), pp. 1278–1293. DOI: `10.1121/1.4863196`.

[308] F. Zotter. “Analysis and synthesis of sound-radiation with spherical arrays.” PhD thesis. KU Graz, IEM, 2009.

[309] F. Zotter. “Sampling Strategies for Acoustic Holography/Holophony on the Sphere.” In: *Proceedings of the NAG/DAGA*. Rotterdam, Netherlands, 2009.

[310] E. Zwicker and H. Fastl. *Psychoacoustics - Facts and Models*. Heidelberg, Germany: Springer, 1990. ISBN: 3540231595.

[311] E. Zwicker and R. Feldtkeller. *Das Ohr als Nachrichtenempfänger*. S. Hirzel Verlag Stuttgart, Germany, 1967.

A Room Acoustical Fundamentals

In the following, some room acoustical fundamentals are provided, including the RIR as a suitable tool for the description of room acoustics. In addition, some room acoustical parameters are described which are important in the context of this thesis.

A.1 Room Impulse Response

A room impulse response (RIR) carries the acoustical properties of a room, i.e., the transfer characteristics from a sound source to a microphone [238]. If a room is considered a linear time-invariant (LTI) system, then, in practice, such measures can be derived from the room impulse response (RIR). Figure A.1 shows a measured RIR.

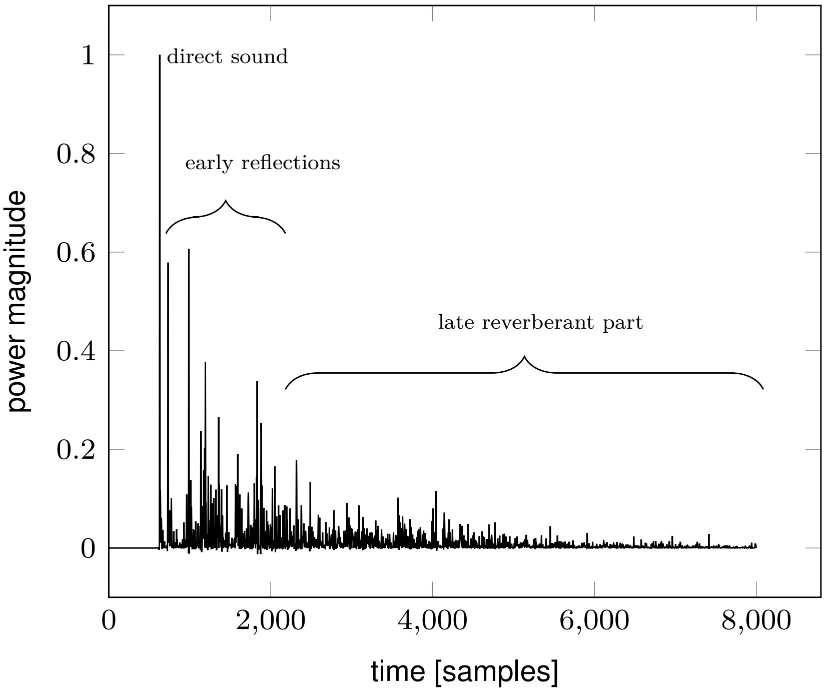


Figure A.1: Measured room impulse response showing the direct sound as well as the early and late part of the sound field.

After the direct sound, the reflections from the surrounding environment arrive successively at the microphone. The RIR is separated into an early and a late sound field, comprising discrete reflections and the late reverberation, respectively [143]. Both influence the perception of spaciousness as will be described with the presentation of common room acoustical (quality) parameters (following [78]). For more information on RIR measurements, the reader is referred to [84, 85].

From the RIR, mainly the direct-to-reverberation ratio was evaluated in early room acoustical quality research, from which basic room acoustic parameters were established. For further information on room acoustics the reader is referred to, for example [143, 186, 187].

A.2 Room Acoustical Parameters

In the following, some important room acoustical parameters are reviewed which are related to the work presented in this thesis.

Reverberation Time RT can be calculated after [83], [174], or [86]. Today, the definition after [245] is common in acoustical measurements, with RT being defined as the time a sound decays by 60 dB after the sound source is turned off. RT can be derived from the impulse response by

$$h^2(t) = \int_t^\infty [h_{SR}(\tau)]^2 d\tau = \int_t^\infty [h_{SR}(\tau)]^2 d(-\tau). \tag{A.1}$$

In practice, the slope between a 5 dB and 35 dB decay is computed, since a 60 dB dynamic range may not be reached due to measurement and background noise.

Sound Strength Studies and listening experiments established the sound strength G as a predictor for subjective loudness ([69], [10] and [261]). In [124], G is investigated for concert hall acoustics and the role between subjective and objective measures is established for the early sound field. Experiments in [21] highlighted G, together with RT, as being an underestimated quality measure in concert hall acoustics. It is shown, that G strongly relates to LEV and ASW.

Clarity The clarity index C_{t_c} describes how well certain components of a signal can be perceived due to blurring from the late reverberation. It describes the applicability of a room for speech and music performances, whereas for speech $t_c = 50\,\mathrm{ms}$ and $t_c = 80\,\mathrm{ms}$ for music. The clarity index C_{t_c} (in dB) is defined after [2]

$$C_{t_c} = 10 \log_{10} \left(\frac{\int_0^{t_c} h_{SR}^2(t) dt}{\int_{t_c}^{\infty} h_{SR}^2(t) dt} \right) \tag{A.2}$$

Another common measure for clarity is the center time T_S [142].

B Spherical Sound Field Special Functions

This section provides the fundamentals for spherical sound field descriptions. Based on the derivation of the wave equation in spherical coordinates, spherical base solutions are presented, i.e., the angular and radial solutions. Then the spherical fourier transform is presented as well as the spherical convolution and correlation.

B.1 Wave Equation in Spherical Coordinates

Fundamental to acoustic wave field formulations in spherical coordinates is the acoustic wave equation in Cartesian coordinates. The homogeneous acoustic wave equation satisfies [218]

$$\nabla_x^2 p(\boldsymbol{x}, t) - \frac{1}{c^2}\frac{\delta^2}{\delta t^2} p(\boldsymbol{x}, t) = 0, \tag{B.1}$$

with $p(\boldsymbol{x}, t)$ denoting the sound pressure measured at position $\boldsymbol{x} = (x, y, z) \in \mathbb{R}^3$ in meters. Variables c and t represent speed of sound (in air under normal ambient conditions) and time in seconds, respectively. $\nabla_{\boldsymbol{x}}^2$ is the Laplacian which for function $f(x, y, z)$ is defined as $\Delta_x^2 f = \delta^2/\delta x^2 f + \delta^2/\delta y^2 f + \delta^2/\delta z^2 f$ in Cartesian coordinates [218].

Assuming a single-frequency sound field, the sound pressure can be expressed by [218]

$$p(\boldsymbol{x}, t) = p(\boldsymbol{x} e^{i\omega t}), \tag{B.2}$$

with ω representing the frequency. The wave equation can be transformed into the Helmholtz equation [218]

$$\nabla_{\boldsymbol{x}}^2 p(k, \boldsymbol{x}) + k^2 p(k, \boldsymbol{x}) = 0, \tag{B.3}$$

with $k = \omega/c$ denoting the wave number in radians per meter. For a broadband plane wave, the pressure field can be written as [218]

$$p(\boldsymbol{x}, t) = Ae^{-i\boldsymbol{k}\cdot\boldsymbol{x}}e^{i\omega t}, \tag{B.4}$$

with A denoting the amplitude and $\boldsymbol{k} \equiv (k_x, k_y, k_z)$ the wave vector representing the wave propagation direction. $\boldsymbol{k} \cdot \boldsymbol{x}$ is the dot-product of vectors $\boldsymbol{k}$ and $\boldsymbol{x}$. Note that the direction of arrival, in contrast to the propagation direction of a wave, is denoted by $\tilde{\boldsymbol{k}} = -\boldsymbol{k}$.

The spherical coordinate system and its relation to Cartesian coordinates is shown in Figure B.1, with $\boldsymbol{r} = (r, \theta, \phi)$.

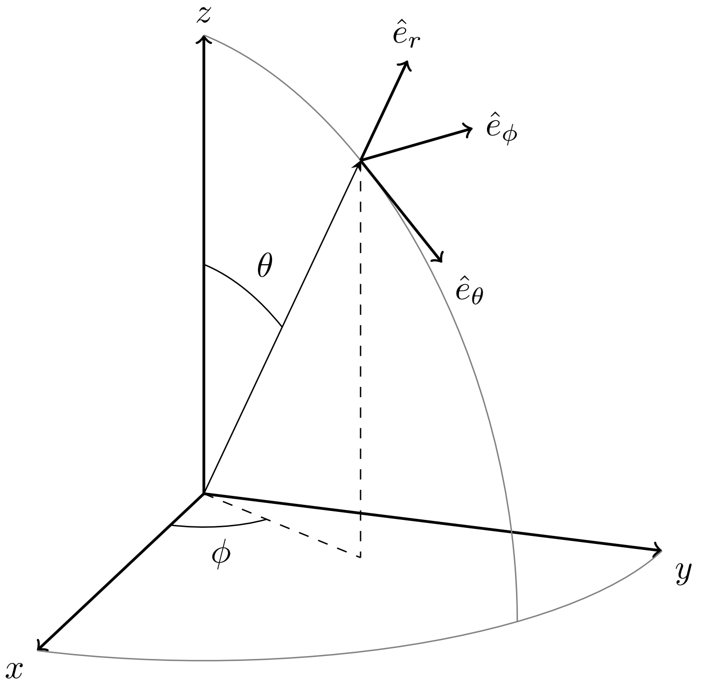


Figure B.1: Schematic of the spherical coordinate system relative to Cartesian coordinates (after [299])

Here, $r = \sqrt{x^2 + y^2 + z^2}$ is the radius, $\theta = \arctan(\sqrt{x^2 + y^2}/z)$ the angle in elevation direction, and $\phi = \arctan(y/x)$ the angle in azimuth direction, with values ranging from $0 \leq \phi \leq 2\pi$, $0 \leq \theta \leq \pi$ and $0 \leq r \leq \infty$, respectively.

The wave equation in spherical coordinates reads [218]

$$\nabla^2_{\boldsymbol{r}} p(\boldsymbol{r}, t) - \frac{1}{c^2}\frac{\delta^2}{\delta t^2} p(\boldsymbol{r}, t) = 0 \tag{B.5}$$

and can be transformed into the Helmholtz equation in spherical coordinates for a single-frequency sound field

$$\nabla_{\boldsymbol{r}}^2 p(k,\boldsymbol{r}) + k^2 p(k,\boldsymbol{r}) = 0, \tag{B.6}$$

with the Laplacian in spherical coordinates which is defined as

$$\nabla_{\boldsymbol{r}}^2 f \equiv \frac{1}{r^2}\frac{\delta}{\delta r}\left(r^2 \frac{\delta}{\delta r} f\right) + \frac{1}{r^2 \sin\theta}\frac{\delta}{\delta\theta}\left(\sin\theta \frac{\delta}{\delta\theta} f\right) + \frac{1}{r^2 \sin^2\theta}\frac{\delta^2}{\delta\phi^2} f. \tag{B.7}$$

Accordingly, the pressure amplitude can be represented by

$$p(\boldsymbol{r},t) = p(\boldsymbol{r})e^{i\omega t}. \tag{B.8}$$

Solution to the Helmholtz Equation in Spherical Coordinates

The solution to the Helmholtz equation is given by separation of variables [299]:

$$p(r,\theta,\phi,t) = R(r)\Theta(\theta)\Phi(\phi)T(t), \tag{B.9}$$

where p can be decomposed into four partial differential equations.

The first

$$\frac{d^2T}{dt^2} + \omega^2 T = 0, \tag{B.10}$$

represents the time dependence with the solution

$$T(t) = e^{i\omega t}, \qquad \omega \in \mathbb{R}. \tag{B.11}$$

The second differential equation, which is depending on ϕ, can be isolated substituting Eq. (B.9) in Eq. (B.6) and multiplying $r^2 \sin^2\theta/p$. It satisfies

$$\frac{d^2\Phi}{d\phi^2} + m^2\Phi = 0, \tag{B.12}$$

with its solution

$$\Phi(\phi) = e^{im\phi}, \qquad m \in \mathbb{Z}, \tag{B.13}$$

and $\phi \in [0, 2\pi)$. The third θ-dependent term can be isolated substituting Eq. (B.13) back into the Helmholtz equation, satisfying

$$\frac{d}{d\mu}\left[(1-\mu^2)\frac{d}{d\mu}\Phi\right] + \left[n(n+1) - \frac{m^2}{1-\mu^2}\right]\Phi = 0, \tag{B.14}$$

with $\mu = \cos\theta$. This so-called associated Legendre differential equation has two solutions, one at $\mu = 1$ and a second, the associated Legendre function of the first kind, which reads

$$\Phi(\theta) = P_n^m(\cos\theta), \qquad n \in \mathbb{N}, \quad m \in \mathbb{Z}. \tag{B.15}$$

The last differential equation relates to r. It can be derived substituting Eq. (B.14) into the Helmholtz equation. After [218], it satisfies

$$\rho^2 \frac{d^2}{d\rho^2} V + 2\rho \frac{d}{d\rho} V + \left[\rho^2 - n(n+1)\right] V = 0, \tag{B.16}$$

with $\rho \equiv kr$ and $V(\rho) \equiv R(r)$. Solutions to this so-called spherical Bessel equation are the spherical Bessel and Hankel functions of the first kind, $j_n(kr)$ and $h_n(kr)$, respectively.

Based on these four solutions over r, θ, ϕ, and t, the solution for the spherical wave equation in spherical coordinates can be written as

$$p(\boldsymbol{r}, t) = j_n(kr) Y_n^m(\theta, \phi) e^{i\omega t} \tag{B.17}$$

or

$$p(\boldsymbol{r}, t) = h_n(kr) Y_n^m(\theta, \phi) e^{i\omega t} \tag{B.18}$$

For different values of n and m, both Eqs. (B.17) and (B.18), or a combination of those solutions, form the basis for sound field descriptions in spherical coordinates. Specific solutions, like sound fields originating from point or plane sources, are given in Chapter 3.1.2.

B.2 Spherical Base Solutions

This section adresses the base solutions for the spherical wave equation comprising angular and radial solutions. The angular solutions presented are the Legendre polynomials and the associated Legendre functions as well as the spherical harmonics; the radial solutions comprise the spherical Bessel, Neumann, and Hankel functions.

Angular Solutions

The angular solutions are given by the following angular functions, i.e., the Legendre Polynomials, the associated Legendre Functions, and the spherical harmonics, which are described next.

Legendre Polynomials

The Legendre polynomials of order n have the general form [299]:

$$\begin{aligned}P_n(x) = \frac{(2n-1)!!}{n!}\Big[&x^n - \frac{n(n-1)}{2\cdot(2n-1)}x^{n-2} \\ &+ \frac{n(n-1)(n-2)(n-3)}{2\cdot 4\cdot(2n-1)(2n-3)}x^{n-4} \\ &- \frac{n(n-1)(n-2)(n-3)(n-4)(n-5)}{2\cdot 4\cdot 6\cdot(2n-1)(2n-3)(2n-5)}x^{n-6} + \dots\Big], \text{ n= 0,1,2,...}\end{aligned} \tag{B.19}$$

with $(2n-1)!! \equiv (2n-1)(2n-3)\cdots 1$. From Eq. (B.19) it follows that [299]

$$P_n(-x) = (-1)^n P_n(x). \tag{B.20}$$

Eq. (B.19) can be rewritten in a more concise way as [299]

$$P_n(x) = \frac{1}{2^n!}\frac{d^n}{dx^n}(x^2-1)^n, \tag{B.21}$$

which is called Rodrigues' Formula. Figure B.2 shows the first five Legendre polynomials for orders $n = 0\dots 4$.

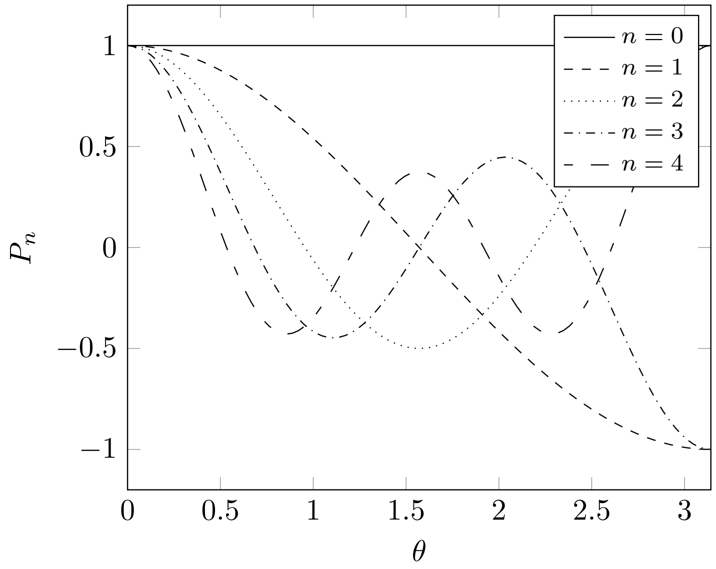


Figure B.2: Legendre polynomials $P_n(cos(\theta))$ for orders $n = 0\dots 4$ as a function of θ

Associated Legendre Functions

The associated Legendre functions are given by $P_n^m(x)$ which, for $m > 0$, is related to the Legendre polynomials by the formula [299]

$$P_n^m(x) = (-1)^m (1 - x^2)^{m/2} \frac{d^m}{dx^m} P_n(x). \tag{B.22}$$

The series representation is described by [299]

$$\begin{aligned} P_n^m(x) = &\frac{(-1)^m (2n-1)!!}{(n-m)!}(1-x^2)^{m/2}\Big[x^{n-m} - \frac{(n-m)(n-m-1)}{2(2n-1)}x^{n-m-2} \\ &+ \frac{(n-m)(n-m-1)(n-m-2)(n-m-3)}{2 \cdot 4(2n-1)(2n-3)}x^{n-m-4} - \cdots\Big]. \end{aligned} \tag{B.23}$$

Fig. B.3 shows the associated Legendre functions $P_n^m(x)$ for $m = 0$ (a) and $m = 1$ (b) for orders $n = 0 \dots 2$ and $n = 4 \dots 6$, respectively.

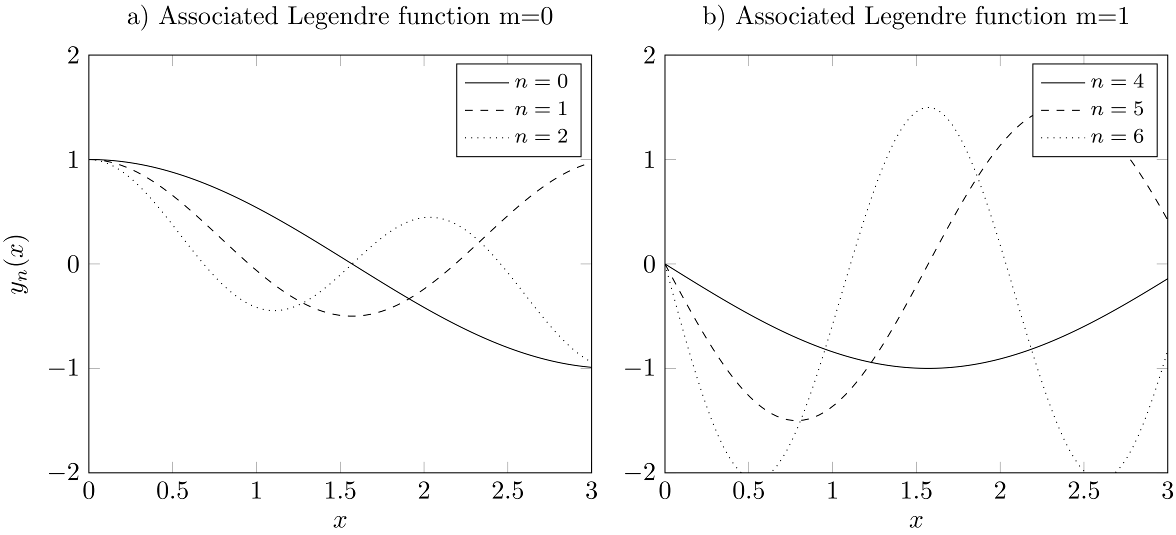


Figure B.3: Associated Legendre functions $P_n^m(x)$ for $m = 0$ in plot a) and $m = 1$ in plot b) for orders $n = 0 \dots 2$

Spherical Harmonics

The spherical harmonics $Y_n^m(\theta, \phi)$ of order n and mode m are base functions used to decompose a distribution on a two-dimensional sphere. They are defined by [299]

$$Y_n^m(\theta, \phi) = \sqrt{\frac{(2n+1)}{4\pi} \frac{(n-m)!}{(n+m)!}} P_n^m(\cos\theta) e^{jm\phi}, \tag{B.24}$$

with P_n^m denoting the associated Legendre function of the first kind. Figure 3.1 shows a set of spherical harmonics for orders $n = 0 \ldots 2$. In the following, some important properties of the spherical harmonics are presented. For more details, see [218].

- *Complex conjugate:* Indicated by the exponential term $e^{jm\phi}$, the spherical harmonics form a set of complex functions. The complex conjugate is written as [218]

$$[Y_n^m(\theta, \phi)]^* = (-1)^m Y_n^m(\theta, \phi). \tag{B.25}$$

- *Limit on degree value:* Spherical harmonics with a degree m higher than the order n are zero, as described by [218]

$$Y_n^m(\theta, \phi) = 0 \quad \forall |m| > n. \tag{B.26}$$

- *Zeros:* Through the associated Legendre function, the spherical harmonics contain $\sin^{|m|\theta}$ terms which define their zeros for $m \neq 0$ [218].

$$Y_n^m(0, \phi) = Y_n^m(\pi, \phi) = 0 \quad \forall m \neq n. \tag{B.27}$$

- *Symmetry:* The spherical harmonics are mirror symmetric with respect to θ and ϕ. This behavior is given by Eqs. (B.28) and (B.29), respectively [218].

$$Y_n^m(\pi - \theta, \phi) = (-1)^{n+m} Y_n^m(\theta, \phi). \tag{B.28}$$

$$Y_n^m(\theta, \phi + \pi) = (-1)^m Y_n^m(\theta, \phi). \tag{B.29}$$

  Another symmetry is along ϕ, relative to the x-axis, i.e., $Y_n^m(\theta, -\phi) = [Y_n^m(\theta, \phi)]$.

- *Opposite direction:* Spherical harmonics at the opposite direction to (θ, ϕ), i.e.,

$(\pi - \theta, \phi + \pi)$, can be written as

$$Y_n^m(\pi - \theta, \phi + \pi) = (-1)^n Y_n^m(\theta, \phi). \tag{B.30}$$

- *Periodicity:* Spherical harmonics are periodic with respect to ϕ with a period of $2\pi/m$, as described by

$$Y_n^m(\theta, \phi + 2\pi/m) = Y_n^m(\theta, \phi). \tag{B.31}$$

- *Orthogonality:* The spherical harmonics are orthogonal over the sphere surface which is described by

$$\int_0^{2\pi} \int_0^{\pi} [Y_n^m(\theta, \phi)]^* \, Y_{n'}^{m'}(\theta, \phi) \sin\theta d\theta d\phi = \delta_{nn'} \delta_{mm'} \tag{B.32}$$

  with $\delta_{nn'} = 1$ for $n = n'$, and zero otherwise.

- *Completeness:* Completeness of spherical harmonics states that

$$\sum_{n=0}^{\infty} \sum_{m=-n}^{n} [Y_n^m(\theta, \phi)]^* \, Y_n^m(\theta', \phi') = \delta(\cos\theta - \cos\theta')\delta(\phi - \phi'). \tag{B.33}$$

- *Addition theorem:* Completeness of spherical harmonics states that

$$\sum_{m=-n}^{n} [Y_n^m(\theta, \phi)]^* \, Y_n^m(\theta', \phi') = \frac{2n+1}{4\pi} P_n(\cos\Theta). \tag{B.34}$$

Radial Solutions

The radial solutions described in the following comprise the spherical Bessel, Neumann, and Hankel functions.

Spherical Bessel, Neumann, and Hankel Functions

Spherical Bessel, Neumann, and Hankel functions are the radial solutions to the Helmholtz equation in spherical coordinates, with the Neumann and Hankel function being the Bessel function of the second and third kind, respectively. Their derivatives are necessary to calculate the radial velocity $v_r(r)$. As described in [308], the spherical Bessel function of the first kind $j_n(kr)$ reads

$$j_n(kr) = (-1)^n (kr)^n \left(\frac{d}{kr\ d(kr)}\right)^n \frac{\sin(kr)}{kr}, \tag{B.35}$$

while the spherical Neumann function $y_n(kr)$ is similarly described by

$$y_n(kr) = (-1)^{n+1} (kr)^n \left(\frac{d}{kr\ d(kr)}\right)^n \frac{\cos(kr)}{kr}. \tag{B.36}$$

Following [299], their derivatives can be written as

$$j'_n(kr) = j_{n-1}(kr) - \frac{n+1}{kr} j_n(kr). \tag{B.37}$$

and

$$y'_n(kr) = y_{n-1}(kr) - \frac{n+1}{kr} y_n(kr). \tag{B.38}$$

The spherical Bessel function of the third kind, i.e., the Hankel function $h_n^{(1)}(kr)$ of the first kind, is a composite of $j_n(kr)$ and $y_n(kr)$:

$$h_n^{(1)}(kr) = j_n(kr) + i y_n(kr), \tag{B.39}$$

with $h_n^{(2)}(kr)$ being the complex conjugate, which is defined as $h_n^{(2)} = {h_n^{(1)}}^*$. The derivative of the Hankel function reads

$$h_n'^{(1,2)}(kr) = h_{n-1}^{(1,2)}(kr) - \frac{n+1}{kr} h_n^{(1,2)}(kr). \tag{B.40}$$

For orders $n = 0 \dots 4$, Figure B.4 shows the Bessel and Neumann functions in plots a) and c), and their derivatives in plots b) and d) respectively.

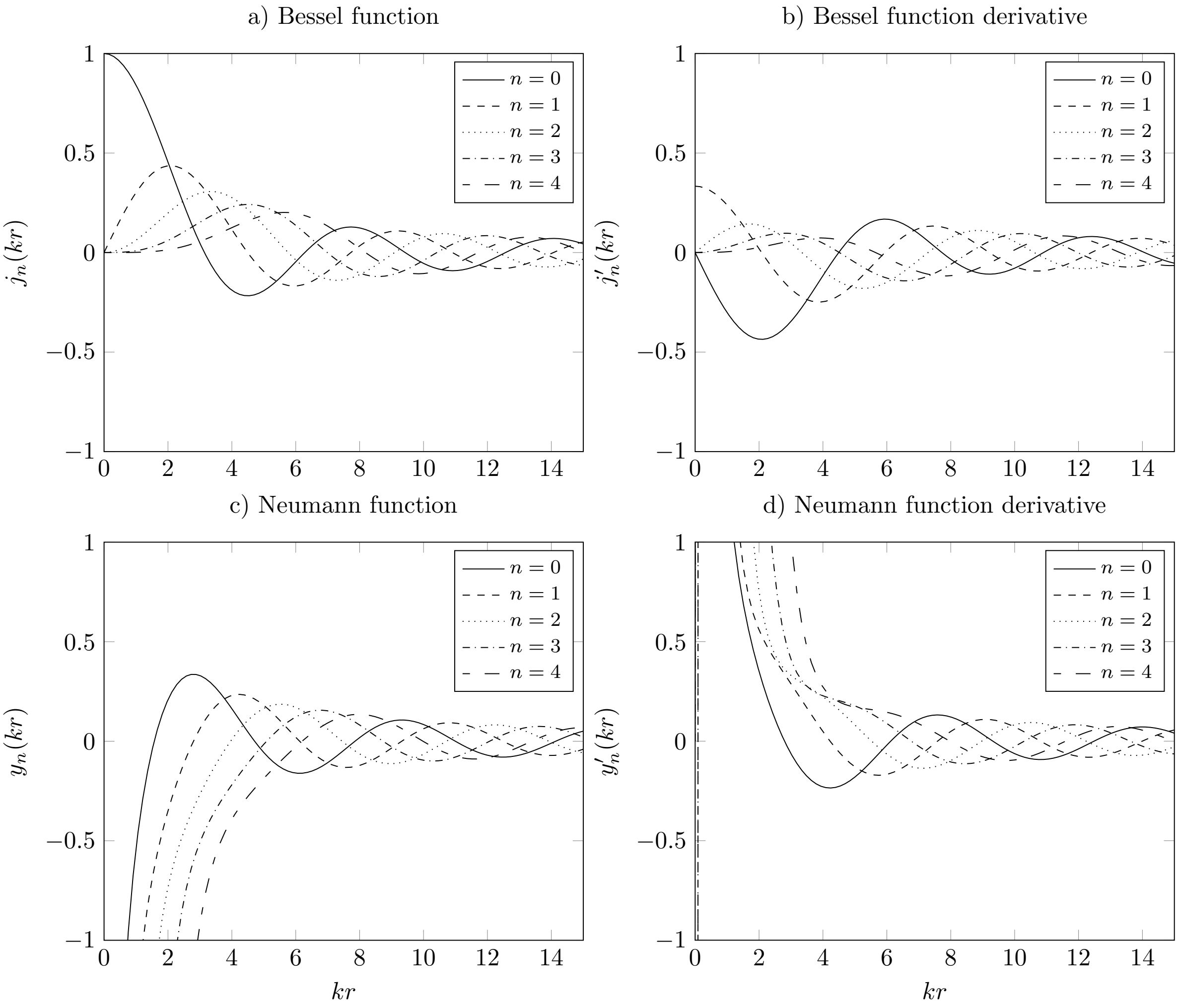


Figure B.4: Spherical Bessel and Neumann functions with their derivatives for orders $n = 0 \dots 4$

B.3 Spherical Fourier Transform

As a basic tool for sound field analysis in spherical coordinates, the SFT and some important properties are described in the following (after [218]). The SFT of a square-integrable function $f(\theta, \phi)$ and its ISFT were described by Eqs. (3.2) and (3.3) resulting in the spherical Fourier coefficients f_{nm}.

- *Parseval's relation* directly follows from the orthogonality and completeness properties of the spherical harmonics, as described by

$$\int_0^{2\pi} \int_0^{\pi} f(\theta, \phi) \left[g(\theta, \phi)\right]^* \sin\theta d\theta d\phi = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} f_{nm} g_{nm}^*. \tag{B.41}$$

- *Linearity* describes the property that scaling and addition of two functions is also applied to their transforms, as written in

$$\begin{aligned} h(\theta,\phi) &= \alpha f(\theta,\phi) + \beta g(\theta,\phi) \\ h_{nm} &= \alpha f_{nm} + \beta g_{nm}, \quad \alpha,\beta \in \mathbb{R}. \end{aligned} \tag{B.42}$$

- *Complex conjugate* of $f(\theta,\phi)$ and its transform reads

$$\begin{aligned} g(\theta,\phi) &= [f(\theta,\phi)]^* \\ g_{nm} &= (-1)^m f^*_{n(-m)}. \end{aligned} \tag{B.43}$$

- *Constancy:* After [218], the SFT is constant along θ and ϕ, i.e., $f(\theta,\phi) = f(\theta)$ and $f(\theta,\phi) = f(\phi)$, which reduce the spherical harmonics coefficients to $f_{nm} = \sqrt{\frac{4\pi}{2n+1}} f_n \delta_{n0}$ and $f_{nm} = f_m C_n^m$, respectively.
- *Symmetry:* The SFT is symmetric with respect to ϕ for a symmetric function, i.e., $f(\theta,\phi) = f(\theta,\pi-\phi)$. The transform which leads to $f_{nm} = f_{n(-m)}$ reads

$$f(\theta,\phi) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} f_{n(-m)} Y_n^m(\theta,\phi). \tag{B.44}$$

- *Sifting:* The sifting property holds for functions on the sphere which are multiplied by a Dirac delta function, as written in

$$\int_0^{2\pi} \int_0^{\pi} f(\theta,\phi)\delta(\cos\theta - \cos\theta^{'})\delta(\phi-\phi^{'})\sin\theta d\theta d\phi = f(\theta^{'},\phi^{'}). \tag{B.45}$$

Discrete SFT

The discrete versions of the SFT and ISFT, denoted DSFT and DISFT, respectively, can both be formulated in matrix notation, following [218]

$$\begin{aligned} \boldsymbol{f_{nm}} &= \boldsymbol{Y}^{\dagger}\boldsymbol{f} \\ \boldsymbol{f} &= \boldsymbol{Y}\boldsymbol{f_{nm}}, \end{aligned} \tag{B.46}$$

with $\boldsymbol{Y}^{\dagger} = (\boldsymbol{Y}^H\boldsymbol{Y})^{-1}\boldsymbol{Y}^H$ being the pseudo-inverse. For a function $f(\theta_q,\phi_q)$ which is sampled with Q samples, the column vector $\boldsymbol{f}$ of length Q is defined as

$$\boldsymbol{f} = [f(\theta_1,\phi_1), f(\theta_2,\phi_2), ..., f(\theta_Q,\phi_Q)]^T, \tag{B.47}$$

and

$$\boldsymbol{f}_{nm} = [f_{00}, f_{1(-1)}, f_{10}, f_{11}, ..., f_{QQ})]^T, \tag{B.48}$$

of length $(N+1)^2$. The spherical harmonics matrix $\boldsymbol{Y}$ of dimension Q $(N+1)^2$ is given by

$$\boldsymbol{Y} = \begin{bmatrix} Y_0^0(\theta_1,\phi_1) & Y_1^{-1}(\theta_1,\phi_1) & Y_1^0(\theta_1,\phi_1) & Y_1^1(\theta_1,\phi_1) & \dots & Y_N^N(\theta_1,\phi_1) \\ Y_0^0(\theta_2,\phi_2) & Y_1^{-1}(\theta_2,\phi_2) & Y_1^0(\theta_2,\phi_2) & Y_1^1(\theta_2,\phi_2) & \dots & Y_N^N(\theta_2,\phi_2) \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_0^0(\theta_Q,\phi_Q) & Y_1^{-1}(\theta_Q,\phi_Q) & Y_1^0(\theta_Q,\phi_Q) & Y_1^1(\theta_Q,\phi_Q) & \dots & Y_N^N(\theta_Q,\phi_Q) \end{bmatrix}. \tag{B.49}$$

Note, that in order to compute $\boldsymbol{f_{nm}}$ matrix $\boldsymbol{Y}$ must be invertible. This is the case when oversampling is employed, such that $Q > (N+1)^2$. For the now over-determined linear system of equations, i.e., $\boldsymbol{f} = \boldsymbol{Y}\boldsymbol{f_{nm}}$, the solution is given in a least-square sense using the pseudo-inverse $\boldsymbol{f_{nm}} = \boldsymbol{Y}^\dagger \boldsymbol{f}$.

For the three sampling schemes, as described in Section 3.1.3, the DSFT can be written by defining the sampling matrix $\boldsymbol{S}$ as

$$\boldsymbol{f}_{nm} = \boldsymbol{S}\boldsymbol{f}, \tag{B.50}$$

resulting in a general sampling formulation $\boldsymbol{S} = \boldsymbol{Y}^\dagger$ [218].

In case of equal-angle and Gaussian sampling, $\boldsymbol{S}$ reads

$$\boldsymbol{S} = \boldsymbol{Y}^H \mathrm{diag}(\alpha), \tag{B.51}$$

with diag(α) being the closed-form expression for the sampling weights $\boldsymbol{\alpha} = [\alpha_0, \alpha_1, ..., \alpha_Q]^T$. Here, no need for matrix inversion is required using $\boldsymbol{f_{nm}} = \boldsymbol{Y}^H \mathrm{diag}(\alpha)\boldsymbol{f}$. The (nearly) uniform sampling schemes can be written by

$$\boldsymbol{S} = \frac{4\pi}{Q}\boldsymbol{Y}^H. \tag{B.52}$$

B.4 Spherical Convolution and Correlation

Following [218], this section describes convolution, which is denoted by $(*)$, and correlation of functions defined over the unit-sphere.

Spherical Convolution

The convolution $y(\mu)$ of two functions, $f(\mu)$ and $g(\mu)$, is defined as

$$y(\mu) = f(\mu) * g(\mu) \quad = \int_{SO(3)} f(\boldsymbol{R}(\xi)\eta)\Lambda(\xi)g(\boldsymbol{R}^{-1}(\xi)\mu)d\xi, \tag{B.53}$$

with $\mu \equiv \mu(\theta, \phi) \in S^2$, $\boldsymbol{R} \equiv \boldsymbol{R}_x(\alpha)\boldsymbol{R}_y(\beta)\boldsymbol{R}_z(\gamma)$ representing a rotation by $\xi \equiv \xi(\alpha, \beta, \gamma) \in SO(3)$, and $\eta = [0, 0, 1]^T$ being the north pole in Cartesian coordinates. For more information on rotation of functions in spherical coordinates, please refer to [218].

Similar to a linear Fourier transform, spherical convolution is equal to multiplication in the spherical harmonics domain, such that

$$y_{nm} = 2\pi\sqrt{\frac{4\pi}{2n+1}} f_{nm} g_{n0}. \tag{B.54}$$

Spherical Correlation

Correlation measures the similarity of two functions. For functions $f(\mu)$ and $g(\mu)$, it is defined as

$$c(\xi) = \int_{S_2} f(\mu) \left[\Lambda(\xi)g(\mu)\right]^* d\mu, \tag{B.55}$$

with ξ denoting the rotation. Based on spherical harmonics, the correlation operation $c(\xi)$ can be written as

$$c(\xi) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \sum_{m'=-n}^{n} f_{nm} g^*_{nm'} \left[D^n_{mm'}(\xi)\right]^*. \tag{B.56}$$

C Tables

Experiment II - Room Simulation Data

Table C.1: Frequency dependent scattering coefficients, fixed over all room simulations.

Boundary	$f = 125$ Hz	$f = 250$ Hz	$f = 500$ Hz	$f = 1000$ Hz	$f = 2000$ Hz	$f = 4000$ Hz
Front Wall	0.5	0.5	0.5	0.5	0.5	0.5
Back Wall	0.5	0.5	0.5	0.5	0.5	0.5
Left Wall	0.5	0.5	0.5	0.5	0.5	0.5
Right Wall	0.5	0.5	0.5	0.5	0.5	0.5
Floor	0.5	0.5	0.5	0.5	0.5	0.5
Ceiling	0.5	0.5	0.5	0.5	0.5	0.5

Table C.2: Frequency dependent absorption coefficients for low reverberation.

Boundary	$f = 125$ Hz	$f = 250$ Hz	$f = 500$ Hz	$f = 1000$ Hz	$f = 2000$ Hz	$f = 4000$ Hz
Front Wall	0.85	0.85	0.85	0.85	0.85	0.85
Back Wall	0.85	0.85	0.85	0.85	0.85	0.85
Left Wall	0.85	0.85	0.85	0.85	0.85	0.85
Right Wall	0.85	0.85	0.85	0.85	0.85	0.85
Floor	0.85	0.85	0.85	0.85	0.85	0.85
Ceiling	0.85	0.85	0.85	0.85	0.85	0.85

Table C.3: Frequency dependent absorption coefficients for high reverberation.

Boundary	$f = 125$ Hz	$f = 250$ Hz	$f = 500$ Hz	$f = 1000$ Hz	$f = 2000$ Hz	$f = 4000$ Hz
Front Wall	0.15	0.15	0.15	0.15	0.15	0.15
Back Wall	0.15	0.15	0.15	0.15	0.15	0.15
Left Wall	0.15	0.15	0.15	0.15	0.15	0.15
Right Wall	0.15	0.15	0.15	0.15	0.15	0.15
Floor	0.15	0.15	0.15	0.15	0.15	0.15
Ceiling	0.15	0.15	0.15	0.15	0.15	0.15

PCA Tables

Table C.4: PCA loadings of first five PCs of predictor prescreening for all errors.

MOV	PC1	PC2	PC3	PC4	PC5
V1	-0.04	-0.07	0.00	-0.14	-0.15
V2	-0.04	-0.07	0.00	-0.13	-0.15
V3	-0.11	-0.32	-0.29	-0.84	0.11
V4	0.00	0.00	0.00	0.00	-0.01
V5	-0.78	0.09	0.59	-0.10	0.10
V6	-0.50	0.49	-0.70	0.10	-0.10
V7	0.00	-0.10	0.02	0.09	-0.44
V8	0.00	0.00	0.00	0.00	0.00
V9	0.00	0.00	0.00	0.00	0.00
V10	0.00	0.00	0.00	0.00	-0.01
V11	-0.01	0.00	0.00	0.01	-0.02
V12	-0.32	-0.77	-0.26	0.40	0.18
V13	-0.02	-0.03	0.00	-0.01	-0.07
V14	-0.07	-0.07	0.06	-0.21	-0.54
V15	0.01	-0.09	0.02	0.08	-0.48
V16	0.00	0.00	0.00	0.00	0.00
V17	0.00	0.00	0.00	0.00	0.00
V18	0.00	0.00	0.00	0.00	0.00
V19	0.00	0.00	0.00	0.00	0.00
V20	-0.05	-0.06	0.02	0.02	-0.18
V21	-0.07	-0.07	0.02	0.08	-0.23
V22	-0.02	-0.03	0.01	0.01	-0.09
V23	-0.04	-0.05	0.00	0.03	-0.16
V24	0.00	0.00	0.00	0.00	0.00
V25	0.00	0.00	0.00	0.00	0.00
V26	0.00	0.00	0.00	0.00	0.00
V27	0.00	0.00	0.00	0.00	0.00
V28	0.00	0.00	0.00	0.00	0.00
V29	0.00	0.00	0.00	0.00	0.00
V30	0.00	0.00	0.00	0.00	0.00
V31	0.00	0.00	0.00	0.00	0.00
V32	-0.03	-0.09	-0.04	0.03	-0.16
V33	-0.08	-0.03	0.00	0.05	-0.06
V34	-0.01	-0.04	-0.02	0.01	-0.07
V35	-0.03	-0.05	-0.03	0.02	-0.11
V36	0.00	0.00	0.00	0.00	0.00
V37	0.00	0.00	0.00	0.00	0.00
V38	0.00	0.00	0.00	0.00	0.00
V39	0.00	0.00	0.00	0.00	0.00

Table C.5: PCA loadings of first seven PCs of predictor prescreening for spatial aliasing.

MOV	PC1	PC2	PC3	PC4	PC5	PC6	PC7
V1	0.02	-0.05	-0.09	0.01	0.13	-0.04	0.09
V2	0.02	-0.05	-0.11	0.03	0.13	-0.04	0.10
V3	0.06	-0.37	0.09	0.54	-0.12	0.72	-0.05
V4	0.00	0.00	0.00	-0.01	0.03	0.01	0.00
V5	0.68	0.18	-0.11	0.04	0.01	-0.04	-0.67
V6	0.68	0.17	0.14	-0.06	-0.05	0.11	0.67
V7	-0.02	-0.09	-0.21	-0.48	-0.21	0.31	0.01
V8	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V9	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V10	0.00	0.00	-0.01	0.00	0.02	0.00	0.00
V11	0.00	-0.01	0.01	-0.05	0.00	-0.01	-0.03
V12	0.23	-0.85	0.26	-0.19	0.06	-0.33	-0.03
V13	0.01	-0.02	-0.04	-0.03	0.12	0.01	0.04
V14	0.03	-0.01	-0.04	-0.26	0.81	0.34	-0.10
V15	-0.02	-0.09	-0.20	-0.50	-0.14	0.32	0.02
V16	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V17	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V18	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V19	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V20	0.02	-0.08	-0.30	0.15	0.18	-0.12	0.16
V21	0.07	-0.10	-0.41	-0.03	-0.24	-0.04	-0.02
V22	0.01	-0.04	-0.13	0.08	0.10	-0.03	0.07
V23	0.03	-0.09	-0.35	0.15	0.09	-0.04	0.04
V24	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V25	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V26	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V27	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V28	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V29	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V30	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V31	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V32	0.02	-0.08	-0.30	0.15	0.18	-0.12	0.16
V33	0.07	-0.10	-0.41	-0.03	-0.24	-0.04	-0.02
V34	0.01	-0.04	-0.13	0.08	0.10	-0.03	0.07
V35	0.03	-0.09	-0.35	0.15	0.09	-0.04	0.04
V36	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V37	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V38	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V39	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table C.6: PCA loadings of first seven PCs of predictor prescreening for measurement noise.

MOV	PC1	PC2	PC3	PC4	PC5	PC6	PC7
V1	-0.05	-0.06	-0.02	-0.05	-0.24	0.14	0.07
V2	-0.05	-0.05	-0.01	-0.05	-0.23	0.13	0.07
V3	-0.15	-0.34	-0.42	-0.81	0.06	-0.10	0.02
V4	0.00	0.00	0.00	0.00	-0.02	0.01	0.00
V5	-0.78	0.19	0.54	-0.21	0.07	0.00	-0.03
V6	-0.42	0.53	-0.70	0.22	-0.05	-0.08	-0.07
V7	-0.01	-0.11	0.05	0.06	-0.30	-0.51	-0.20
V8	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V9	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V10	0.00	0.00	0.00	0.00	-0.01	-0.01	0.00
V11	0.00	0.00	0.00	0.01	-0.01	-0.01	0.02
V12	-0.41	-0.72	-0.18	0.46	0.18	0.13	-0.14
V13	-0.02	-0.02	0.00	0.01	-0.10	0.04	0.00
V14	-0.08	-0.04	0.00	-0.06	-0.75	0.45	-0.12
V15	0.00	-0.10	0.05	0.05	-0.34	-0.48	-0.31
V16	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V17	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V18	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V19	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V20	-0.05	-0.05	0.00	0.05	-0.13	-0.17	0.21
V21	-0.07	-0.06	0.05	0.05	-0.09	-0.39	0.33
V22	-0.02	-0.02	0.00	0.02	-0.07	-0.09	0.08
V23	-0.04	-0.04	0.01	0.03	-0.09	-0.21	0.22
V24	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V25	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V26	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V27	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V28	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V29	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V30	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V31	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V32	-0.03	-0.07	-0.05	0.09	-0.16	0.05	0.51
V33	-0.08	0.00	0.00	0.04	0.04	-0.03	0.37
V34	-0.02	-0.04	-0.02	0.04	-0.08	0.03	0.25
V35	-0.04	-0.04	-0.03	0.05	-0.07	0.00	0.38
V36	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V37	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V38	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V39	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table C.7: PCA loadings of first seven PCs of predictor prescreening for microphone position errors.

MOV	PC1	PC2	PC3	PC4	PC5	PC6	PC7
V1	0.04	-0.08	0.03	-0.22	-0.06	-0.11	0.13
V2	0.04	-0.08	0.03	-0.21	-0.05	-0.11	0.15
V3	0.10	-0.31	0.33	-0.74	0.26	0.35	-0.03
V4	0.00	-0.01	0.00	-0.01	-0.01	-0.01	0.01
V5	0.79	0.09	-0.56	-0.11	0.16	0.03	-0.10
V6	0.50	0.46	0.70	0.13	-0.19	0.03	0.02
V7	0.00	-0.10	-0.09	0.00	-0.47	0.39	-0.11
V8	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V9	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V10	0.00	0.00	0.00	0.00	-0.01	0.00	0.00
V11	0.01	0.01	-0.01	0.01	-0.04	0.07	-0.04
V12	0.30	-0.79	0.24	0.43	0.04	-0.11	-0.16
V13	0.01	-0.03	0.01	-0.04	-0.05	-0.07	0.04
V14	0.08	-0.09	-0.03	-0.38	-0.52	-0.65	-0.11
V15	-0.01	-0.09	-0.09	-0.01	-0.53	0.35	-0.21
V16	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V17	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V18	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V19	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V20	0.05	-0.06	-0.03	0.02	-0.16	0.07	0.27
V21	0.07	-0.07	-0.09	0.04	-0.18	0.28	0.44
V22	0.02	-0.03	-0.02	0.01	-0.08	0.04	0.12
V23	0.04	-0.05	-0.04	0.02	-0.11	0.14	0.31
V24	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V25	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V26	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V27	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V28	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V29	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V30	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V31	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V32	0.04	-0.10	0.01	0.00	-0.06	-0.18	0.42
V33	0.08	-0.02	-0.01	0.05	0.07	0.00	0.38
V34	0.02	-0.05	0.01	0.00	-0.03	-0.09	0.21
V35	0.04	-0.06	0.00	0.01	0.00	-0.06	0.34
V36	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V37	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V38	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V39	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Tables for System Error Prediction in Free-field Sound Fields

Table C.8: R^2 and RMSE for estimated performance for spatial aliasing regression in free-field environments, model parameters, and preprocessing steps.

Model	R^2	RMSE	Optimal tuning parameter	Data preprocessing
LM	0.83	14.64	-	BoxCox, C&S
RLM	0.84	14.49	-	BoxCox, C&S
GLM	0.82	14.68	-	BoxCox, C&S, PCA
PLS	0.86	14.60	ncomp = 5	BoxCox, C&S, PCA
ANN	0.86	12.26	size = 21, decay = 0.1	BoxCox, C&S
MARS	0.87	14.69	degree = 1, nprune = 7	BoxCox, C&S
SVM$_{lin}$	0.86	13.62	cost = 0.25	BoxCox, C&S, PCA
SVM$_{rad}$	0.81	13.51	cost = 64, sigma = 0.06357128	BoxCox, C&S, PCA
KNN	0.73	17.25	$K = 5$	BoxCox, C&S, PCA
BT	0.87	12.57	mincriterion = 0.7321053	
RF	0.90	10.55	mtry = 15	
SGB	0.86	11.74	ntrees = 300, depth = 15, shrinkage = 0.1, minobsinnode = 10	

Table C.9: R^2 and RMSE for estimated performance for measurement noise regression in free-field environments, model parameters, and preprocessing steps.

Model	R^2	RMSE	Optimal tuning parameter	Data preprocessing
LM	0.81	15.55	-	BoxCox, C&S
RLM	0.78	16.28	-	BoxCox, C&S
GLM	0.80	15.92	-	BoxCox, C&S, PCA
PLS	0.81	15.49	ncomp = 4	BoxCox, C&S, PCA
ANN	0.86	12.73	size = 39, decay = 0.1	BoxCox, C&S
MARS	0.88	10.57	degree = 1, nprune = 5	BoxCox, C&S
SVM$_{lin}$	0.83	15.95	cost = 8	BoxCox, C&S, PCA
SVM$_{rad}$	0.84	14.07	cost = 64, sigma = 0.1088506	BoxCox, C&S, PCA
KNN	0.85	15.59	$K = 5$	BoxCox, C&S, PCA
BT	0.73	18.15	mincriterion = 0.1131579	
RF	0.84	14.27	mtry = 8	
SGB	0.84	14.13	ntrees = 1000, depth = 9, shrinkage = 0.1, minobsinnode = 10	

Table C.10: R^2 and RMSE for estimated performance for positioning error regression in free-field environments, model parameters, and preprocessing steps.

Model	R^2	RMSE	Optimal tuning parameter	Data preprocessing
LM	0.67	19.07	-	BoxCox, C&S
RLM	0.69	19.17	-	BoxCox, C&S
GLM	0.64	19.59	-	BoxCox, C&S, PCA
PLS	0.74	18.50	ncomp = 4	BoxCox, C&S, PCA
ANN	0.69	17.58	size = 27, decay = 0.1	BoxCox, C&S
MARS	0.72	16.62	degree = 2, nprune = 3	BoxCox, C&S
SVM_{lin}	0.71	18.47	cost = 16	BoxCox, C&S, PCA
SVM_{rad}	0.70	18.04	cost = 4, sigma = 0.1299385	BoxCox, C&S, PCA
KNN	0.60	20.41	$K = 7$	BoxCox, C&S, PCA
BT	0.47	23.42	mincriterion = 0.1131579	
RF	0.73	16.93	mtry = 15	
SGB	0.76	15.80	ntrees = 550, depth = 17, shrinkage = 0.1, minobsinnode = 10	

Tables for System Error Prediction in Reflective Sound Fields

Table C.11: Overall accuracy (OA), model parameters, and preprocessing steps for error classification in reverberant environments.

Model	OA	Optimal tuning parameter	Data preprocessing
LDA	0.71	-	BoxCox, C&S
PLS	0.69	ncomp = 19	BoxCox, C&S
RDA	0.69	gamma = 0.5, lambda = 0.5	BoxCox, C&S, PCA
ANN	0.81	size = 21, decay = 0.1	BoxCox, C&S
FDA	0.67	degree = 4, nprune = 30	BoxCox, C&S
SVM_{lin}	0.68	cost = 1	BoxCox, C&S, PCA
SVM_{rad}	0.73	cost = 256, sigma = 0.1	BoxCox, C&S, PCA
KNN	0.69	$K = 17$	BoxCox, C&S, PCA
BT	0.77	model = rule, win = false, trials = 20	
RF	0.78	mtry = 20	
SGB	0.77	ntrees = 150, depth = 3, shrinkage = 0.1, minobsinnode = 10	

Table C.12: R^2 and RMSE for estimated performance for spatial aliasing regression in reverberant environments, model parameters, and preprocessing steps.

Model	R^2	RMSE	Optimal tuning parameter	Data preprocessing
LM	0.79	10.39	-	BoxCox, C&S
RLM	0.78	10.29	-	BoxCox, C&S
GLM	0.78	10.30	-	BoxCox, C&S, PCA
PLS	0.86	8.72	ncomp = 19	BoxCox, C&S, PCA
ANN	0.75	11.14	size = 17, decay = 0.1	BoxCox, C&S
MARS	0.83	9.75	degree = 1, nprune = 3	BoxCox, C&S
SVM_{lin}	0.79	10.41	cost = 0.25	BoxCox, C&S, PCA
SVM_{rad}	0.75	11.05	cost = 1, sigma = 0.2967131	BoxCox, C&S, PCA
KNN	0.76	11.14	$K = 5$	BoxCox, C&S, PCA
BT	0.77	10.54	mincriterion = 0.8352632	
RF	0.78	10.54	mtry = 39	
SGB	0.76	11.04	ntrees = 150, depth = 1, shrinkage = 0.1, minobsinnode = 10	

Table C.13: R^2 and RMSE for estimated performance for noise error regression in reverberant environments, model parameters, and preprocessing steps.

Model	R^2	RMSE	Optimal tuning parameter	Data preprocessing
LM	0.97	5.74	-	BoxCox, C&S
RLM	0.96	5.86	-	BoxCox, C&S
GLM	0.97	5.76	-	BoxCox, C&S, PCA
PLS	0.98	4.51	ncomp = 14	BoxCox, C&S, PCA
ANN	0.97	4.94	size = 39, decay = 0.1	BoxCox, C&S
MARS	0.98	3.99	degree = 1, nprune = 3	BoxCox, C&S
SVM_{lin}	0.97	5.92	cost = 0.25	BoxCox, C&S, PCA
SVM_{rad}	0.98	4.81	cost = 0.5, sigma = 0.1367072	BoxCox, C&S, PCA
KNN	0.98	4.57	$K = 7$	BoxCox, C&S, PCA
BT	0.98	4.39	mincriterion = 0.99	
RF	0.98	3.95	mtry = 5	
SGB	0.96	6.39	ntrees = 150, depth = 3, shrinkage = 0.1, minobsinnode = 10	

Table C.14: R^2 and RMSE for estimated performance for positioning error regression in reverberant environments, model parameters, and preprocessing steps.

Model	R^2	RMSE	Optimal tuning parameter	Data preprocessing
LM	0.92	9.11	-	BoxCox, C&S
RLM	0.92	9.32	-	BoxCox, C&S
GLM	0.93	9.05	-	BoxCox, C&S, PCA
PLS	0.94	8.12	ncomp = 1	BoxCox, C&S, PCA
ANN	0.92	8.76	size = 39, decay = 10	BoxCox, C&S
MARS	0.93	8.15	degree = 1, nprune = 3	BoxCox, C&S
SVM_{lin}	0.91	9.56	cost = 0.5	BoxCox, C&S, PCA
SVM_{rad}	0.93	9.04	cost = 0.5, sigma = 0.1130316	BoxCox, C&S, PCA
KNN	0.93	8.11	$K = 7$	BoxCox, C&S, PCA
BT	0.92	8.76	mincriterion = 0.5257895	
RF	0.94	7.55	mtry = 19	
SGB	0.88	11.27	ntrees = 50, depth = 2, shrinkage = 0.1, minobsinnode = 10	

Table C.15: Error prediction performances for the test data in reverberant environments, showing R^2 and RMSE for all models.

model	Aliasing		Noise		Pos. error	
	R^2	RMSE	R^2	RMSE	R^2	RMSE
LM	0.65	14.58	0.97	7.22	0.95	7.67
RLM	0.68	14.02	0.97	6.80	0.95	7.44
GLM	0.65	14.58	0.97	7.22	0.95	7.67
PLS	0.82	11.31	0.98	6.30	0.95	7.94
ANN	0.66	14.33	0.96	7.14	0.95	7.77
MARS	0.67	14.13	0.97	5.97	0.94	8.54
SVM_{lin}	0.72	12.97	0.96	6.84	0.95	7.53
SVM_{rad}	0.62	15.28	0.96	6.84	0.94	9.76
KNN	0.60	15.80	0.97	6.03	0.93	9.39
BT	0.73	13.05	0.97	6.34	0.92	9.84
RF	0.87	9.47	0.97	5.74	0.93	9.26
SGB	0.90	8.36	0.95	8.05	0.92	9.92

Table C.16: R^2 and RMSE for estimated performance for aliasing error regression in reverberant environments using models trained with free-field data from Experiment I. Also model tuning parameters and preprocessing steps are given.

Model	R^2	RMSE	Optimal tuning parameter	Data preprocessing
LM	0.83	13.31	-	BoxCox, C&S, PCA
RLM	0.84	14.15	-	BoxCox, C&S, PCA
GLM	0.83	13.37	-	BoxCox, C&S, PCA
PLS	0.88	11.94	ncomp = 5	BoxCox, C&S
ANN	0.85	12.17	size = 39, decay = 10	BoxCox, C&S, PCA
MARS	0.90	11.20	degree = 1, nprune = 10	BoxCox, C&S
SVM_{lin}	0.82	14.38	cost = 0.5	BoxCox, C&S, PCA
SVM_{rad}	0.81	14.00	cost = 8, sigma = 0.07081454	BoxCox, C&S, PCA
KNN	0.71	17.24	$K = 5$	BoxCox, C&S, PCA
BT	0.86	12.44	mincriterion = 0.01	
RF	0.91	9.61	mtry = 19	
SGB	0.92	9.38	ntrees = 300, depth = 15, shrinkage = 0.1, minobsinnode = 10	

Table C.17: R^2 and RMSE for estimated performance for noise error regression in reverberant environments using models trained with free-field data from Experiment I. Also model tuning parameters and preprocessing steps are given.

Model	R^2	RMSE	Optimal tuning parameter	Data preprocessing
LM	0.76	15.16	-	BoxCox, C&S, PCA
RLM	0.77	15.06	-	BoxCox, C&S, PCA
GLM	0.77	14.92	-	BoxCox, C&S, PCA
PLS	0.80	14.50	ncomp = 4	BoxCox, C&S
ANN	0.78	14.32	size = 39, decay = 10	BoxCox, C&S
MARS	0.82	13.57	degree = 2, nprune = 4	BoxCox, C&S
SVM_{lin}	0.77	15.06	cost = 0.25	BoxCox, C&S, PCA
SVM_{rad}	0.80	14.75	cost = 2, sigma = 1.0524	BoxCox, C&S, PCA
KNN	0.83	12.98	$K = 5$	BoxCox, C&S, PCA
BT	0.74	15.84	mincriterion = 0.8352632	
RF	0.83	13.06	mtry = 29	
SGB	0.82	13.40	ntrees = 100, depth = 14, shrinkage = 0.1, minobsinnode = 10	

Table C.18: R^2 and RMSE for estimated performance for positioning error regression in reverberant environments using models trained with free-field data from Experiment I. Also model tuning parameters and preprocessing steps are given.

Model	R^2	RMSE	Optimal tuning parameter	Data preprocessing
LM	0.64	18.66	-	BoxCox, C&S, PCA
RLM	0.63	19.61	-	BoxCox, C&S, PCA
GLM	0.62	18.55	-	BoxCox, C&S, PCA
PLS	0.68	17.28	ncomp = 1	BoxCox, C&S
ANN	0.70	16.94	size = 39, decay = 10	BoxCox, C&S
MARS	0.75	15.05	degree = 1, nprune = 8	BoxCox, C&S
SVM$_{lin}$	0.60	20.03	cost = 3	BoxCox, C&S, PCA
SVM$_{rad}$	0.64	17.98	cost = 2, sigma = 0.5981529	BoxCox, C&S, PCA
KNN	0.68	17.34	$K = 11$	BoxCox, C&S, PCA
BT	0.58	20.22	mincriterion = 0.06157895	
RF	0.77	14.66	mtry = 2	
SGB	0.76	14.66	ntrees = 300, depth = 19, shrinkage = 0.1, minobsinnode = 10	

Table C.19: Error prediction performances for the test data in reverberant environments using models trained with free-field data from Experiment I, showing R^2 and RMSE for all models.

model	Aliasing		Noise		Pos. error	
	R^2	RMSE	R^2	RMSE	R^2	RMSE
LM	0.49	76.06	0.63	37.90	0.80	56.20
RLM	0.48	66.95	0.63	37.65	0.80	57.88
GLM	0.49	76.06	0.63	37.90	0.80	56.20
PLS	0.44	109.12	0.43	57.16	0.89	49.44
ANN	0.82	13.34	0.74	22.33	0.68	19.72
MARS	0.42	98.21	0.30	47.54	0.01	61.98
SVM$_{lin}$	0.47	70.85	0.63	37.86	0.80	61.50
SVM$_{rad}$	0.47	68.72	0.23	33.97	0.06	41.21
KNN	0.86	14.59	0.79	22.25	0.81	21.34
BT	0.31	24.30	0.47	33.74	0.82	20.31
RF	0.46	22.84	0.68	23.84	0.81	21.17
SGB	0.46	27.23	0.78	21.76	0.85	21.38

Tables for ASW and LEV Prediction

Table C.20: R^2 and RMSE for estimated performance for ASW regression, model parameters, and preprocessing steps.

Model	R^2	RMSE	Optimal tuning parameter	Data preprocessing
LM	0.88	11.47	-	BoxCox, C&S
RLM	0.87	12.04	-	BoxCox, C&S
GLM	0.89	11.31	-	BoxCox, C&S, PCA
PLS	0.93	8.81	ncomp = 1	BoxCox, C&S, PCA
ANN	0.93	8.85	size = 39, decay = 10	BoxCox, C&S
MARS	0.93	8.27	degree = 1, nprune = 3	BoxCox, C&S
SVM_{lin}	0.87	12.20	cost = 0.25	BoxCox, C&S, PCA
SVM_{rad}	0.90	10.02	cost = 0.5, sigma = 0.2848997	BoxCox, C&S, PCA
KNN	0.89	10.03	$K = 7$	BoxCox, C&S, PCA
BT	0.92	9.24	mincriterion = 0.6289474	
RF	0.95	6.92	mtry = 29	
SGB	0.95	7.28	ntrees = 1000, depth = 4, shrinkage = 0.1, minobsinnode = 10	

Table C.21: R^2 and RMSE for estimated performance for LEV regression, model parameters, and preprocessing steps.

Model	R^2	RMSE	Optimal tuning parameter	Data preprocessing
LM	0.87	6.02	-	BoxCox, C&S
RLM	0.86	11.47	-	BoxCox, C&S
GLM	0.87	11.28	-	BoxCox, C&S, PCA
PLS	0.92	8.46	ncomp = 5	BoxCox, C&S, PCA
ANN	0.91	9.23	size = 39, decay = 10	BoxCox, C&S
MARS	0.92	8.21	degree = 2, nprune = 6	BoxCox, C&S
SVM_{lin}	0.86	11.96	cost = 0.25	BoxCox, C&S, PCA
SVM_{rad}	0.88	10.73	cost = 1, sigma = 0.2188579	BoxCox, C&S, PCA
KNN	0.90	9.46	$K = 7$	BoxCox, C&S, PCA
BT	0.92	8.45	mincriterion = 0.6805263	
RF	0.93	7.69	mtry = 13	
SGB	0.95	7.15	ntrees = 350, depth = 6, shrinkage = 0.1, minobsinnode = 10	

Tables for Array Configuration and Reflection Classification

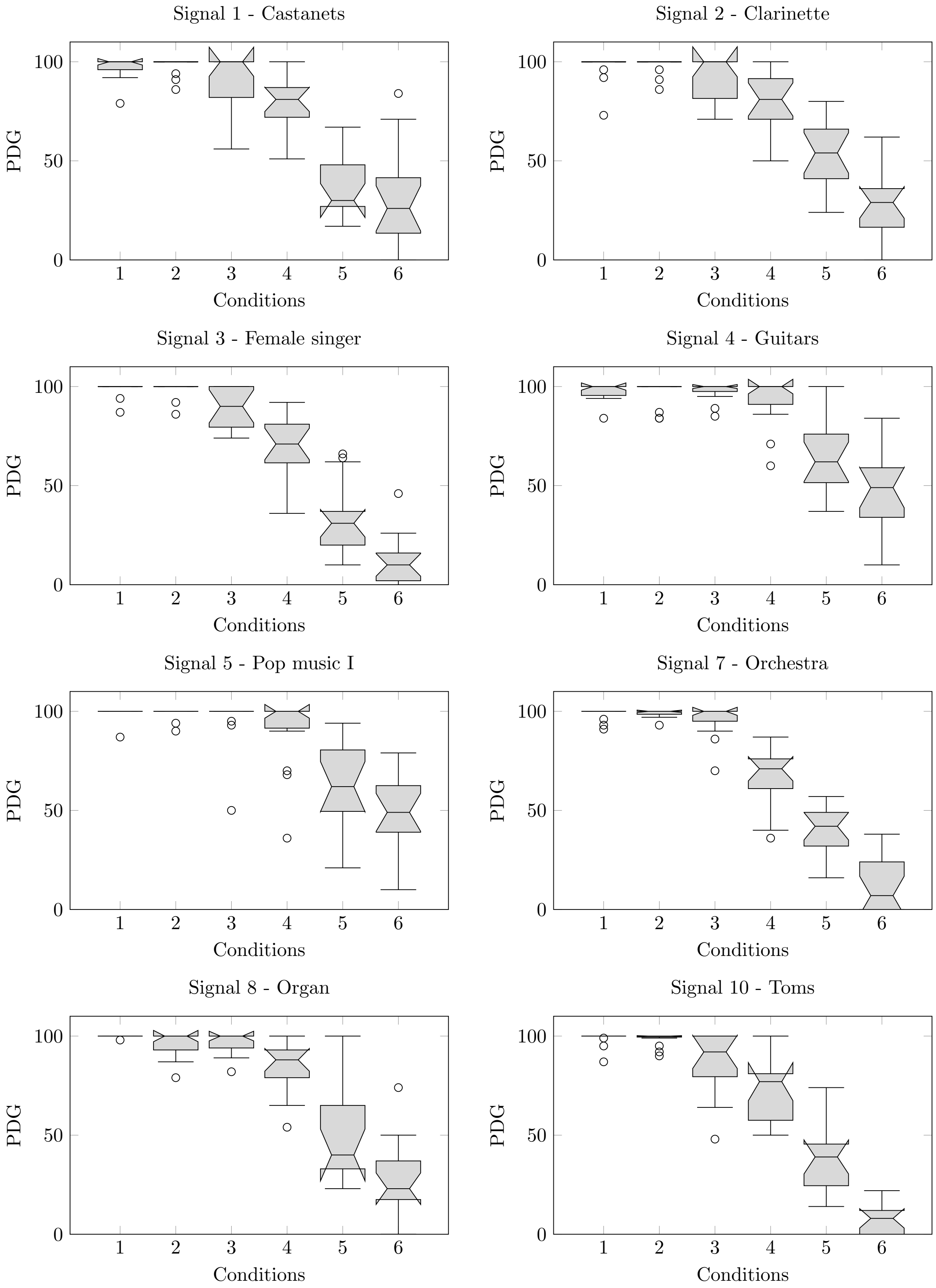
Table C.22: Overall accuracy (OA), model parameters, and preprocessing steps for sound field order classification.

Model	OA	Optimal tuning parameter	Data preprocessing
LDA	0.81	-	BoxCox, C&S
PLS	0.80	ncomp = 32	BoxCox, C&S
RDA	0.66	gamma = 0.5, lambda = 1.5	BoxCox, C&S, PCA
ANN	0.77	size = 3, decay = 0.1	BoxCox, C&S
FDA	0.72	degree = 1, nprune = 17	BoxCox, C&S
SVM$_{lin}$	0.75	cost = 32	BoxCox, C&S, PCA
SVM$_{rad}$	0.78	cost = 32, sigma = 0.06357128	BoxCox, C&S, PCA
KNN	0.71	$K = 9$	BoxCox, C&S, PCA
BT	0.81	mtry = 20	
RF	0.82	mtry = 20	
SGB	0.85	ntrees = 150, depth = 3, shrinkage = 0.1, minobsinnode = 10	

Table C.23: Overall accuracy (OA), model parameters, and preprocessing steps for room classification.

Model	OA	Optimal tuning parameter	Data preprocessing
LDA	0.97	-	BoxCox, C&S
PLS	0.97	ncomp = 33	BoxCox, C&S
RDA	0.96	gamma = 0.2, lambda = 1.5	BoxCox, C&S, PCA
ANN	0.97	size = 3, decay = 0.01	BoxCox, C&S
FDA	0.93	degree = 2, nprune = 10	BoxCox, C&S
SVM$_{lin}$	0.96	cost = 4	BoxCox, C&S, PCA
SVM$_{rad}$	0.96	cost = 4, sigma = 0.06357128	BoxCox, C&S, PCA
KNN	0.96	$K = 5$	BoxCox, C&S, PCA
BT	0.95	mtry = 20	
RF	0.98	mtry = 2	
SGB	0.98	ntrees = 150, depth = 2, shrinkage = 0.1, minobsinnode = 10	

D Perceptual Data


Signal 1 - Castanets

Signal 2 - Clarinette

Signal 3 - Female singer

Signal 4 - Guitars

Signal 5 - Pop music I

Signal 7 - Orchestra

Signal 8 - Organ

Signal 10 - Toms

Figure D.1: PDGs of spatial aliasing in free-fields for signals 1, 2, 3, 4, 5, 7, 8, and 10 shown by notched boxplots. The median is indicated with the black line and outliers with a black circle.

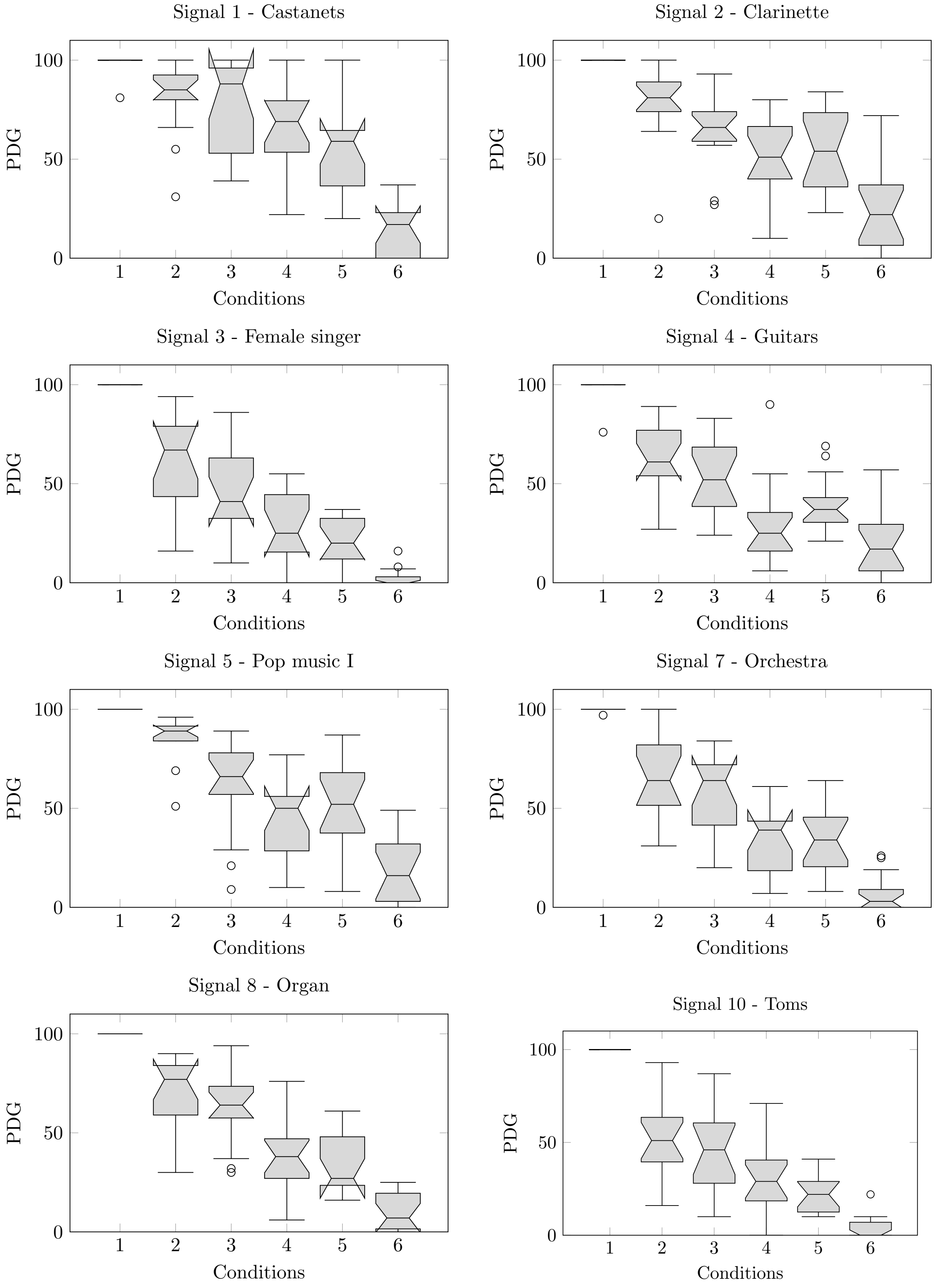


Figure D.2: PDGs of measurement noise error in free-field for signals for signals 1, 2, 3, 4, 5, 7, 8, and 10 shown by notched boxplots. The median is indicated with the black line and outliers with a black circle.

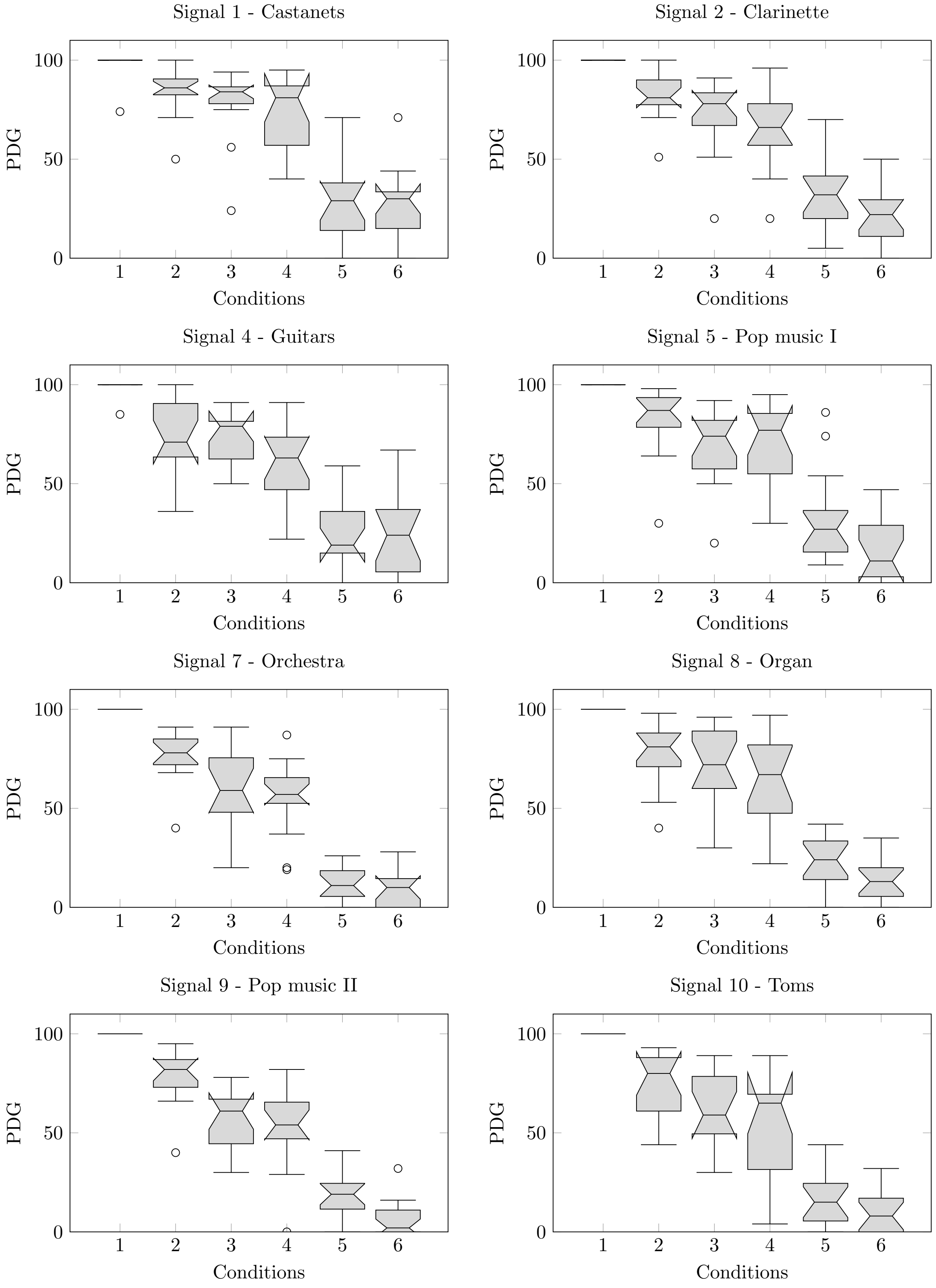


Figure D.3: PDGs of microphone positioning errors in free-fields for signals 1, 2, 4, 5, 7, 8, 9, and 10 shown by notched boxplots. The median is indicated with the black line and outliers with a black circle.

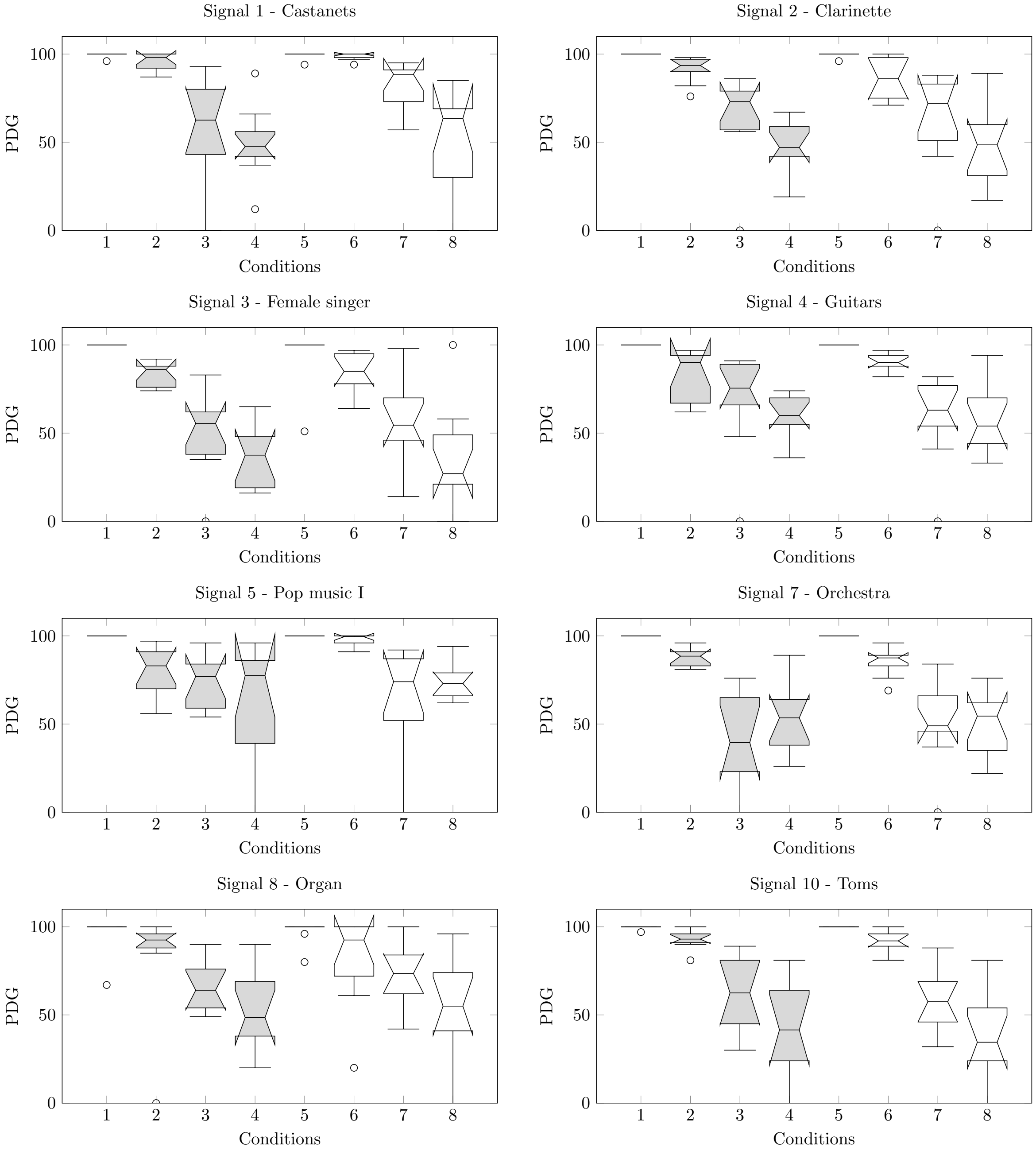


Figure D.4: PDGs of spatial aliasing in simple room geometries for signals 1, 2, 3, 4, 5, 7, 8, and 10 shown by notched boxplots. The median is indicated with the black line and outliers with a black circle.

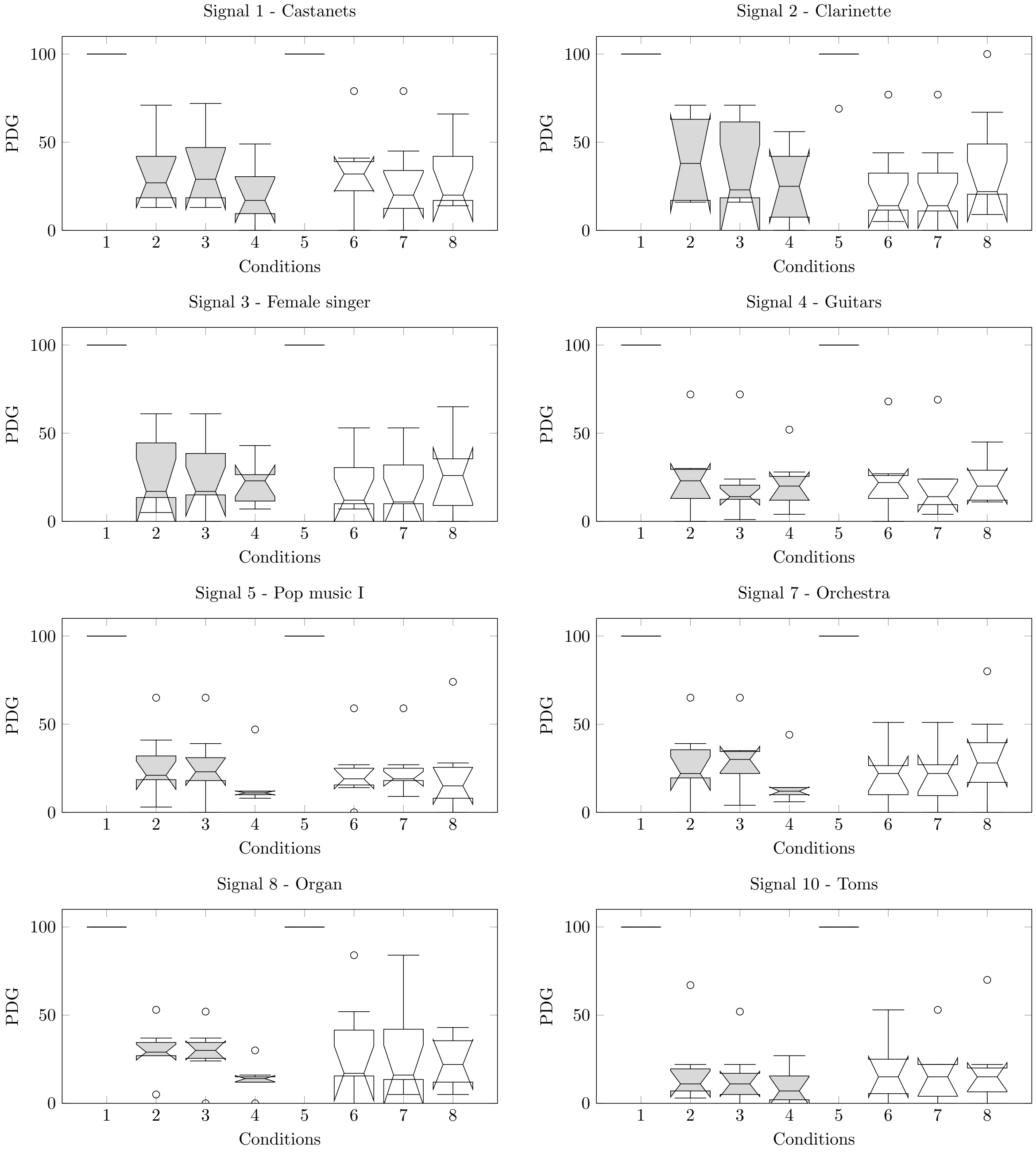


Figure D.5: PDGs of measurement noise in simple room geometries for signals 1, 2, 3, 4, 5, 7, 8, and 10 shown by notched boxplots. The median is indicated with the black line and outliers with a black circle.

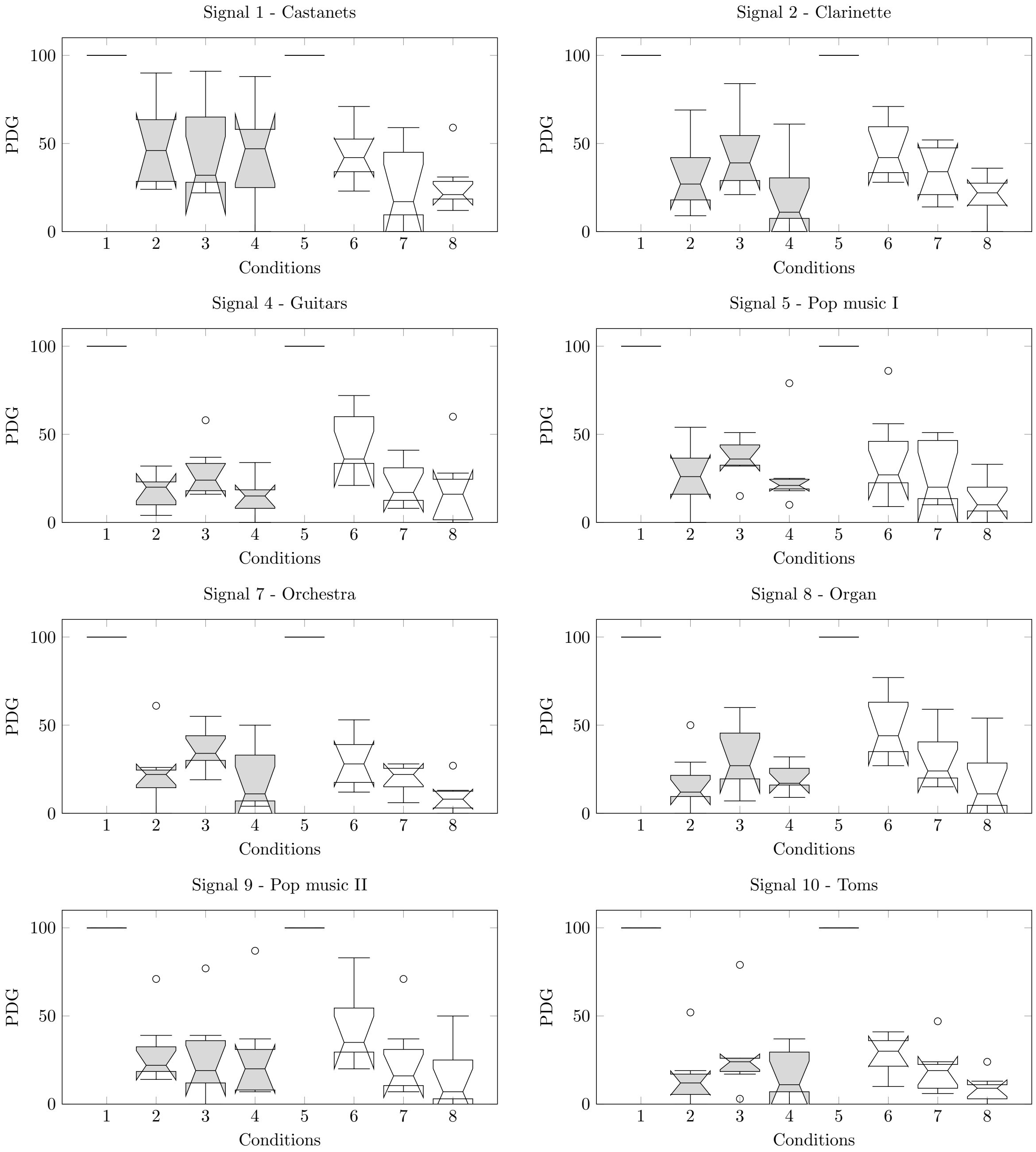
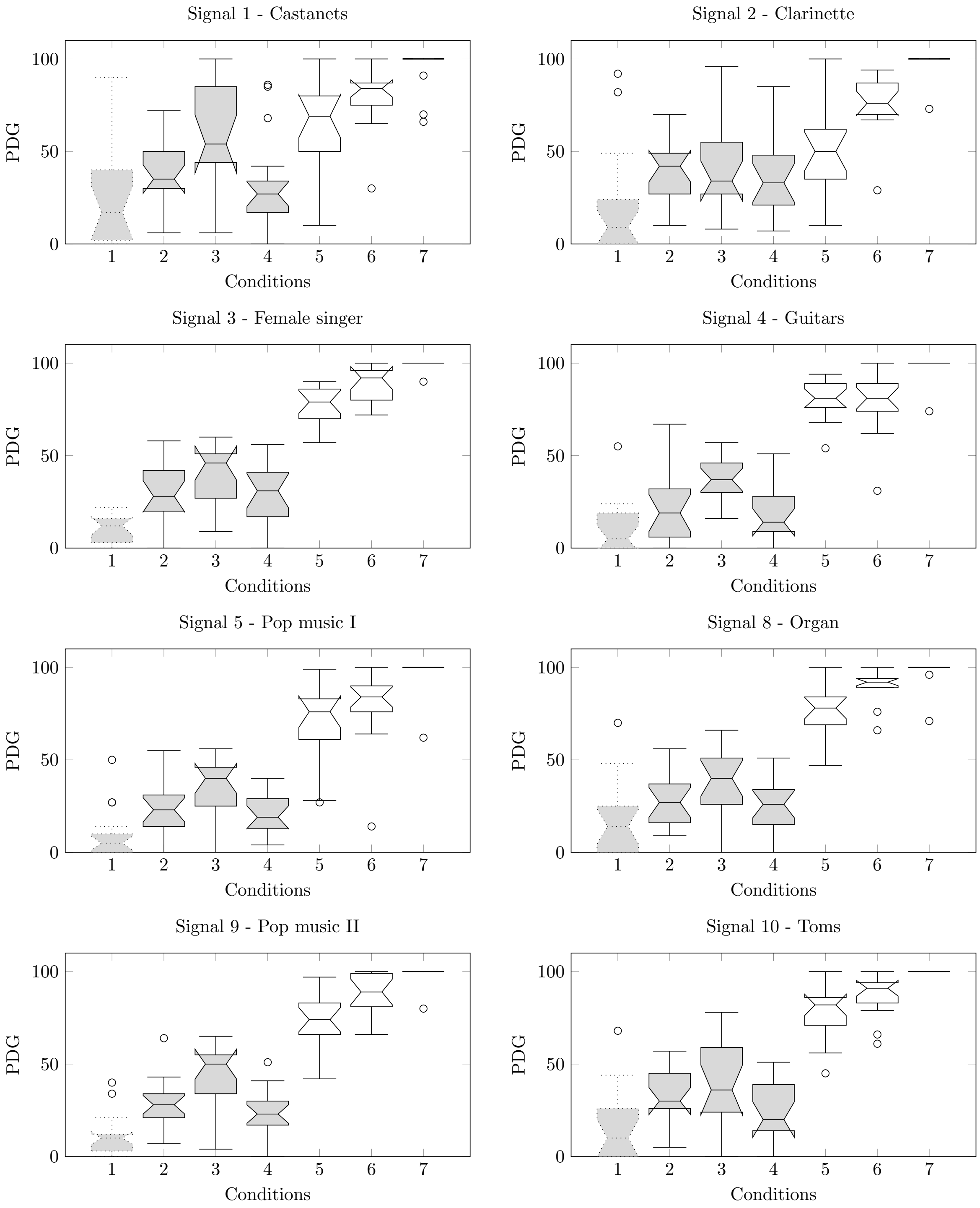


Figure D.6: PDGs of microphone positioning errors in simple room geometries for signals 1, 2, 4, 5, 7, 8, 9, and 10 shown by notched boxplots. The median is indicated with the black line and outliers with a black circle.


Signal 1 - Castanets

Signal 2 - Clarinette

Signal 3 - Female singer

Signal 4 - Guitars

Signal 5 - Pop music I

Signal 8 - Organ

Signal 9 - Pop music II

Signal 10 - Toms

Figure D.7: PDGs for ASW for signals 1, 2, 3, 4, 5, 8, 9, and 10 from Table 4.1 shown by notched boxplots. The median is indicated with the black line and outliers with a black circle.

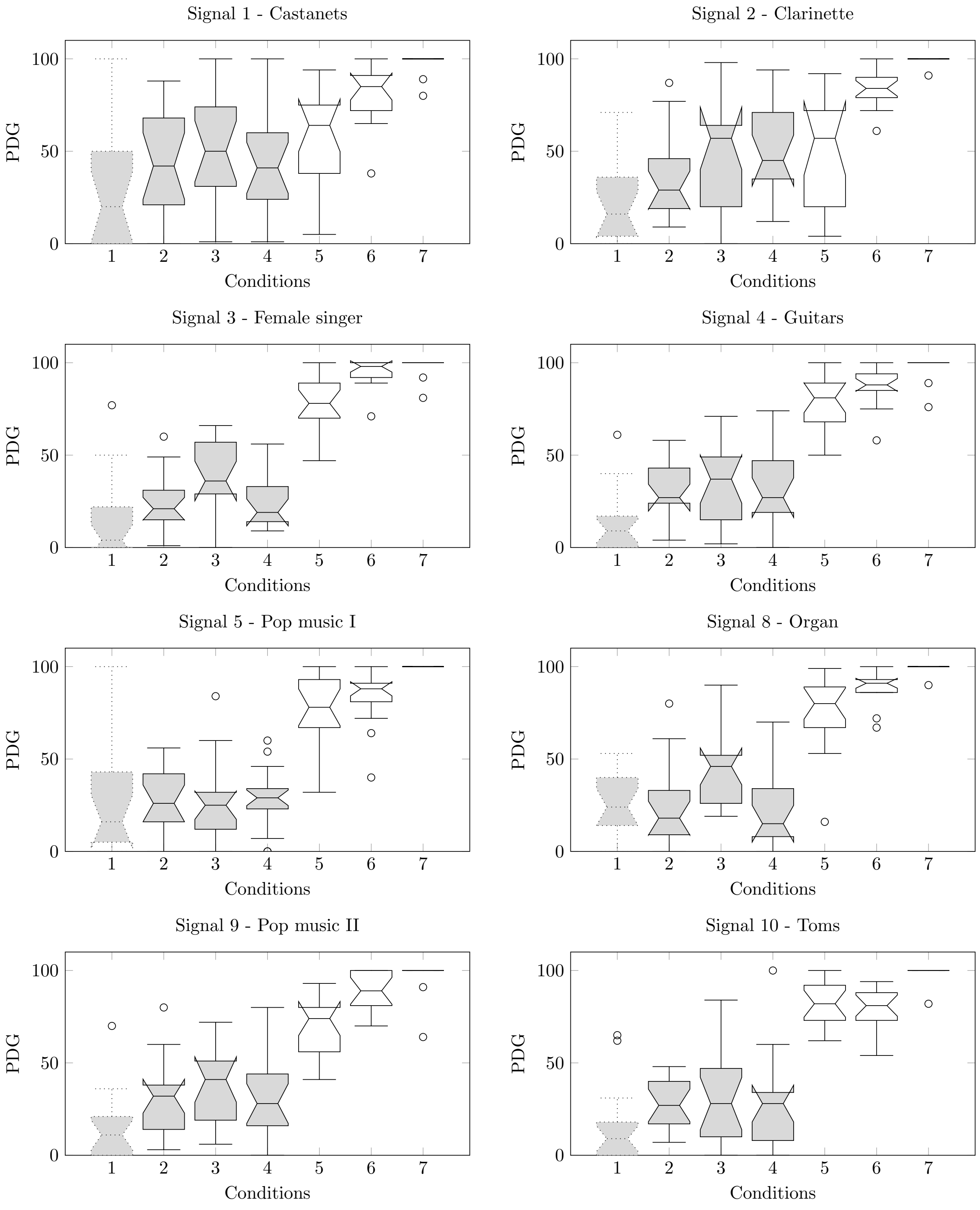


Figure D.8: PDGs for LEV for signals 1, 2, 3, 4, 5, 8, 9, and 10 from Table 4.1 shown by notched boxplots. The median is indicated with the black line and outliers with a black circle.

Eigenständigkeitserklärung

Ich versichere, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen direkt oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet.

Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form einer Prüfungsbehörde vorgelegt.

Ich bin darauf hingewiesen worden, dass die Unrichtigkeit der vorstehenden Erklärung als Täuschungsversuch bewertet wird und gemäß §7 Abs. 10 der Promotionsordnung den Abbruch des Promotionsverfahrens zur Folge hat.

Ilmenau, July 2, 2019