Development and Evaluation of an Immersive Audio Conferencing System

Martin Rothbucher

Technische Universität München
Lehrstuhl für Datenverarbeitung

Development and Evaluation of an Immersive Audio Conferencing System

Martin Rothbucher

Vollständiger Abdruck der von der Fakultät für Elektrotechnik und Informationstechnik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor-Ingenieurs (Dr.-Ing.)

genehmigten Dissertation.

Vorsitzender: Univ.-Prof. Dr.-Ing./Univ. Tokio Martin Buss

Prüfer der Dissertation:

1. Univ.-Prof. Dr.-Ing. Klaus Diepold
2. Univ.-Prof. Dr.-Ing. Georg Färber (em.)

Die Dissertation wurde am 23. April 2014 bei der Technischen Universität München eingereicht und durch die Fakultät für Elektrotechnik und Informationstechnik am 22. Oktober 2014 angenommen.

Martin Rothbucher. *Development and Evaluation of an Immersive Audio Conferencing System.* Dissertation, Technische Universität München, Munich, Germany, 2014.

Institute for Data Processing, Technische Universität München, 80290 München, Germany, http://www.ldv.ei.tum.de.

Abstract

In an audio conferencing situation a remote conferee often faces difficulties distinguishing between the other conference participants. As a consequence conference effectiveness and efficiency are lower than in a conferencing situation where all participants are placed in one room. In order to reduce the disadvantage of being physically not present in the conference room, it would be beneficial to provide the remote conferee with a virtually separated playback of the different conference participants.

In this thesis an immersive audio conferencing system is developed and evaluated which is able to play back the conference contributions to a remote conferee in a spatially separated manner. In order to virtually synthesize the conferees to different positions around the remote conference participant, each participant has to be assigned to an individual transmission channel. The assignment of the active conferee is accomplished via a sound acquisition system that is able to fulfill the task of channel assignment by using sound source localization, sound source separation and online speaker recognition techniques. Due to a required low mouth to ear delay, the assignment algorithm is restricted to a small algorithmic latency.

To further improve the head-related transfer function (HRTF) based immersive playback of the conference contributions at the remote site, different HRTF individualization approaches are applied and acoustic measurement approaches are investigated.

Finally, the developed conferencing system is evaluated by numerous listening tests that are explicitly tailored to the conference situation in order to allow one to draw meaningful conclusions about the conferencing system's sound acquisition algorithm and the different immersive playback modes of the system. Besides the traditional objective measures, e.g. signal to noise ratios, and established listening tests, e.g. localization tests for immersive playback, recent evaluation concepts such as quality of experience and cognitive load are applied. The listening tests demonstrate that the sound acquisition system succeeds to assign the conferees to individual channels within the required small algorithmic latency. Furthermore, the listening evaluation unveils that the efficiency and effectiveness of a conference can be improved significantly by use of the developed conferencing system whereas individual acoustically measured or individualized HRTFs do not improve the subjectively perceived quality of a conference situation compared to a generic set of dummy-head HRTFs.

Contents

1. Introduction **1**
1.1. The Perfect Teleconferencing System 1
1.2. Requirements 3
1.3. State of the Art 4
1.4. Limitations of Existing Teleconferencing Systems 6
1.5. Formulation of the Research Problem 6
1.6. Contributions 9
1.7. Overview 12

I. Sound Acquisition for Immersive Playback **15**

2. Sound Source Localization **19**
2.1. Teleconference Sound Localization 20
2.2. Design of the Array 22
2.3. Sound Source Localization: Experiment 25
2.4. Concluding Remarks: Sound Source Localization 27

3. Sound Source Separation **31**
3.1. Teleconference Sound Source Separation 32
3.2. Sound Source Separation: Experiment 36
3.3. Concluding Remarks: Sound Source Separation 39

4. Online Speaker Recognition **43**
4.1. Teleconference Online Speaker Recognition Algorithm 44
4.2. Online Speaker Recognition: Experiment 46
4.3. Concluding Remarks: Online Speaker Recognition 50

5. Channel Assignment **51**
5.1. Channel Assignment Algorithms 51
5.2. Channel Assignment: Experiment 52
5.3. Concluding Remarks: Channel Assignment 55

II. Immersive Playback 57

6. HRTF-Customization by Selection 61
6.1. Preliminary Listening Experiment 62
6.2. Concluding Remarks: HRTF Customization by Selection 63

7. HRTF Customization by Regression 65
7.1. HRTF Customization 66
7.2. HRTF Customization Methods 67
7.3. HRTF Customization: Experiment 69
7.4. Concluding Remarks: HRTF Customiation by Regression 71

8. HRTF Measurement 73
8.1. HRIR Measurement Approaches 74
8.2. HRTF Measurement: Experiment 75
8.3. Concluding Remarks: HRTF Measurement 80

III. System Evaluation 81

9. System Evaluation Concepts 85
9.1. Sound Localization Performance 85
9.2. Quality of Experience 86
9.3. Cognitive Load 89

10. Listening Experiments 93
10.1. Overall Experimental Settings 93
10.2. Sound Localization Evaluation 96
10.3. Quality of Experience Evaluation 107
10.4. Cognitive Load Evaluation 120

11. Conclusion 127

Bibliography 144

1. Introduction

„*Not seeing separates us from things, not hearing separates us from people.*“

This quote attributed to Kant suggests that the sense of hearing is essential for us to communicate with each other. Recent economic and social developments require more and more interactive communication using teleconference systems to manage work progress and projects whose project partners are spread all over the world. Focusing on a professional sector's project management, multi-party teleconference meetings certainly can not replace a real meeting but teleconferencing systems often enable an information exchange at all, and can therefore have strong influences on the project progress. Nowadays, the different spatially separated project partners usually conduct acoustic or video teleconference meetings to coordinate project progresses. To perform this task, the different groups within a project are often represented in the teleconference by one person or a group of conferees which participate in the conference in front of a computer, by using smart phones or conferencing systems.

1.1. The Perfect Teleconferencing System

The various thinkable teleconferencing situations can be divided into two main groups, illustrated in Figure 1.1. The first group of teleconference situations consists of hands free conferencing of more than one conferee within a conference room that is acoustically equipped with loudspeakers playing back the conference contributions of remote participants and a microphone system which is shared by the conferees in the conference room. The second group of teleconference situations consists of conferees which are equipped with an individual microphone and playback device, e.g. using a head set in combination with a computer or a smart phone. It is also possible for conferencing to use a laptop's microphone for recording of the speech contributions and the laptop's loudspeaker system to playback the conference discussions of the other participants.

A perfect teleconferencing system gives the conference participant the feeling of actually sitting around one conference table with the other remote conferees. In the illustrated conference room situation, the perfect teleconferencing system reliably determines who is speaking when without any delay and the system assigns the microphone array recorded speech contributions of the respective conferee to an individual transmission channel. The microphone array for the sound acquisition can be placed on a table in the conference room and is ready to use out of box without the need of any acoustic calibration. Furthermore, the assignment also works reliably in echoic and noisy environments and the assigned

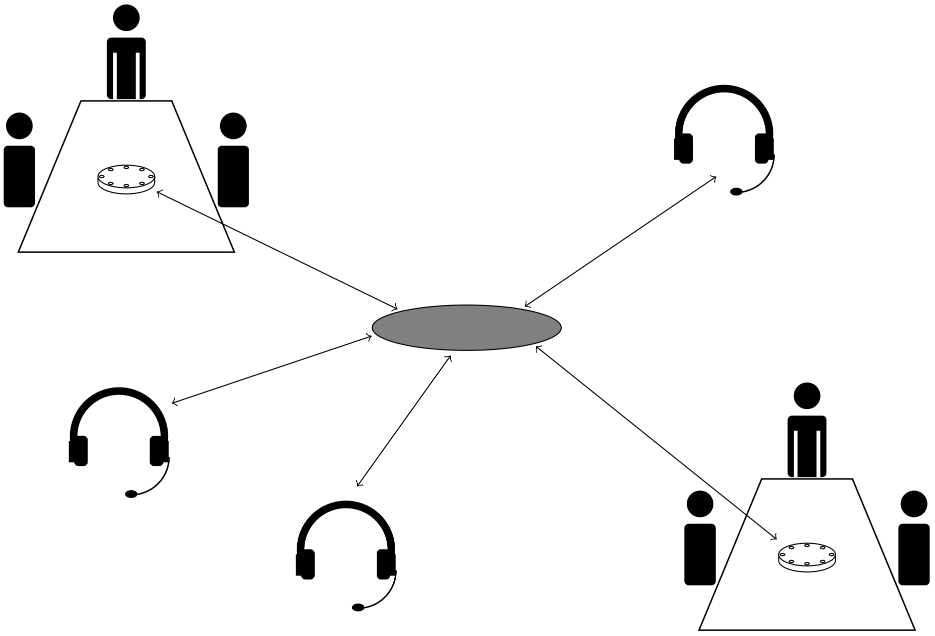

Figure 1.1.: Schematic overview of teleconferencing situations

speech signal is free of echo and noise. Remote conference participants are played back spatially separated to the conferees in the conference room to ease the differentiation of the speech contributions of remote conference participants. Of course, the loudspeaker's playbacks in the conference room do not have disturbing influences at the remote listener's site. The conference room equipment should use standard low-cost hardware which is easy to deinstall, e.g., to be rebuilt in another room.

The perfect conference contribution playback for the second group at the remote site would present the conference contributions without any mouth to ear delay in a spatially distributed manner. The 3D sound synthesis should also incorporate individual differences of the remote conferees with respect to their individual acoustic perception. Thereby however, the users shall be spared from extra individualization effort. Comparable to real hearing habits, movements of the remote participants are captured and the sound is adapted accordingly. The immersive playback is possible with headphones as well as with computer loudspeakers. To achieve the immersive playback for the remote conferee everything needed should be already available at a standard office workplace.

The different tasks to be solved for the described acoustic teleconference system can be summarized to sound acquisition for immersive playback, sound transmission and immersive playback.

Beside audio considerations, visual systems capture each conferee and the associated nonverbal signals which are transmitted and immersively displayed to the other participants.

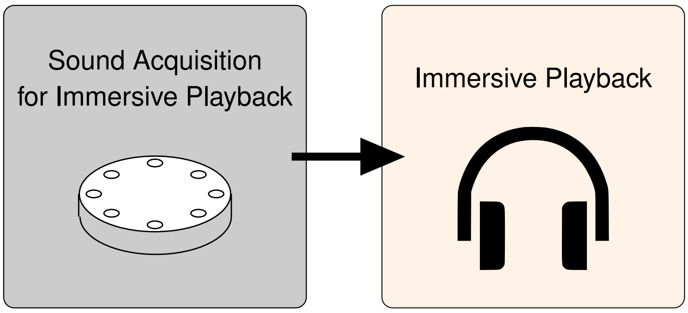

Figure 1.2.: Schematic overview of the thesis focus

In this thesis I will focus on a scenario with one or more remote conference participants equipped with a microphone and stereo headphones. The remote participants conduct an audio teleconference with a group of conferees placed in a conference room that is equipped with a microphone system on a table. Within this scenario I will concentrate on the sound acquisition for immersive playback in the conference room and the immersive playback at the remote site as schematically illustrated in Figure 1.2.

1.2. Requirements

Based on the considerations about the perfect acoustic teleconferencing system and the limitation of the whole system regarding sound acquisition for immersive playback in the conference room and the immersive playback for the remote user, several requirements can be defined:

Sound Acquisition for Immersive Playback

- The sound acquisition system's algorithms are supposed to function in a real world office environment, meaning the presence of reverberation and noise.
- The microphone array is supposed to be placed on a conference table and should therefore be compact in size.
- The channel assignment algorithms should be able to process the conference contributions without extensive calibration or prior knowledge of acoustic transfer functions.
- The assignment of the speakers to their individual channel should be independent from the chronology of the contribution of the respective conference participants.

- The assignment of the active conferees has to be achieved on-line.
- The assignment system is text independent, meaning that no prior knowledge about the conference content is required to fulfill the channel assignment.
- The assignment system adapts to a physical change of a participant's position within the conference room.
- The assignment system only needs speaker dependent information which can be gained in a short introduction round.

Immersive Playback

- The assigned conference participants should be played back at virtually synthesized positions around the remote listener. Each person in the conference room is always synthesized to a fixed position, which enables the listener also to differentiate between the speakers by utilizing the direction of the incoming sound.
- The 3D sound synthesis should be done with an economically justified effort. An additional user expense in terms of user dependent calibration of the system, e.g., by measuring individual head-related transfer functions (HRTFs) is only justified by a leap in quality of the teleconferencing system.
- The immersive playback should be done with standard office equipment. Therefore, no custom built hardware should be utilized.

Complete System

- The ITU-T G.114 recommendation [77] provides information about the acceptable mouth to ear delay of a telecommunication system. According to the recommendation, a user satisfying teleconferencing system should have a mouth to ear delay of less than approximately 275 ms. Therefore, the algorithmic latency for the sound acquisition, the sound transmission and the immersive sound playback has to be considered.

1.3. State of the Art

On the one hand, there has been great progress for conferencing systems concerning the quality of speech by increasing bandwidth of internet connections and by implementing capable algorithms to suppress noise. As a simple conference solution, freeware is often used to conduct software-based teleconferences. One popular teleconferencing freeware is *Skype*, using direct peer-to-peer connection between the users. Audio codecs used by *Skype* seek to improve speech quality but do not use binaural techniques to generate

3D sound. *Ekiga* and *Mumble*, open source VoIP softwares, were equipped with binaural sound rendering to make use of the well-known cocktail party effect [7], enabling the user to virtually differentiate and understand simultaneously talking speakers [70, 136]. Recently, *Symonics* presented the idea of an easy to use server-based 3D teleconferencing system. A similar idea of a server-based 3D sound rendering also was presented in [138]. For the described software-based binaural approaches, a channel assignment is not required since the conferees have their own microphone. Therefore, the individual sound streams can be virtually synthesized at the remote listener's site.

On the other hand, a popular solution to be installed at the conference room is a conferencing telephone that is able to record the participants and transmit the mono recordings to remote conference attendants without any spatial information. One of the market leaders in this category is *Polycom*. Beside standard conferencing telephones, high quality but very expensive solutions for professional teleconferencing are available on the market offered by *Cisco* or *Huawei*. At each conference site, a specially equipped conference room is necessary to record each conferee by an individual camera and microphone. The conferees in one conference room sit on one side of the conference table whereas a screen and a loudspeaker for each remote conference participant "sit" on the other table side. In such a way, a high degree of audio immersion is expected to be achieved, because the audio signals are in fact played back at physically different places in the conference room.

The project 3D VIVANT (3D live immerse video audio interactive multimedia) investigates future 3D audio-visual technologies such as 3D holoscopic content and spatial audio technology to provide the future immersive environment. Within 3D VIVANT, project MARVIN (Microphone Array for Realtime and Versatile Interpolation) can be considered as a promising approach to achieve binaural recordings with a spherical microphone array. The microphone array imitades a head and dependent on the orientation of the listener, a pair of microphones is chosen for playback. The recordings can then be played back by a technology called Binaural Sky [109]. MARVINs main application is supposed to be the recording of sound for immersive 3D video and audio experiences, which is comparable to teleconferencing systems.

In robotic applications, high fidelity telepresence also is investigated. In order to improve interactions between the human (operator) and the robot (teleoperator) in human centered robotic systems, it is important to equip the robotic platform with multimodal human-like sensing, e.g. vision, haptic and audition [86]. The scenario of a robot that is equipped with microphones and an operator who should be capable to pick up the teleoperators auditory scene resembles an acoustic teleconference scenario and should therefore be mentioned in this section.

1.4. Limitations of Existing Teleconferencing Systems

The existing acoustic conferencing solutions do not fully meet the requirements for the defined conference scenario where a group of participants in a conference room are in discussion with one or more remote conferees that are equipped with headphones.

Many VoIP based systems like *Skype* do not offer binaural 3D sound synthesis at all or work with a static set of transfer functions disregarding the users individual acoustic perception that is also influenced by the users unique geometric features (*Ekiga, Mumble, Symonics*). Furthermore these systems do not address the sound acquisition requirements in scenarios where a group of conferees is placed in one conference room.

The widespread conference telephones, e.g. by *Polycom*, do not assign the active conferee to the respective transmission channel. Consequently, a spatial separated playback of the different conference room participants is not possible.

MARVIN does also not assign the conference room participants to individual channels since the spatially separated playback at the remote site is done by direct playback of the binaural recordings of the array. This also makes it difficult to automatically respond to a conferee position change during the teleconference.

The telepresence systems by *Cisco* and *Huawei* are very expensive and require huge installation efforts. Moreover, each remote conference participant has to also be equipped with such a system to benefit from the systems technical capabilities.

The afore mentioned telepresence system for robotic applications [86] can be applied to teleconference situations. Instead of being installed on a teleoperator, the compact sound acquisition module can be installed into a conference room and the immersive playback system at the operators site can be adapted for the remote teleconference participant. However, the system described in [86] is not capable of the channel assignment since the system only works with localization data for direct virtual playback of the robots surrounding sounds.

1.5. Formulation of the Research Problem

Inspired by my work on the project "Acoustic Telepresence: Binaural Directional Hearing and Immersive Audio" within the collaborative research center SFB-453 "High Fidelity Telepresence and Teleaction" and based on the defined requirements of the targeted teleconference scenario I address the following research questions, illustrated in Figure 1.3 at the system level:

- **Sound acquisition**: Is it possible to construct a sound acquisition for the teleconferencing system that reliably assigns each conferee in the echoic conference room to an individual channel and thus enable spatial separated playback of the conference participants at the remote site?

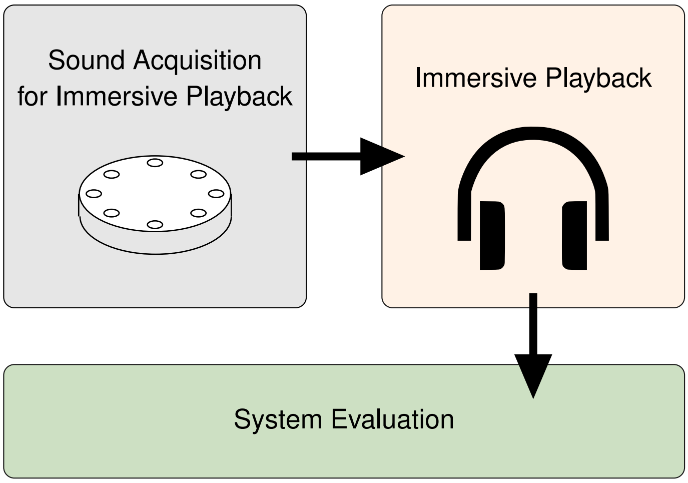

Figure 1.3.: Schematic overview of the research problem

- **Immersive playback**: Is it possible to improve the remote conferee's head-related transfer function (HRTF) based virtual playback of the assigned conference participant's contributions by using individual or customized HRTF datasets with respect to the user?

- **Evaluation**: How can the quality of the developed sound acquisition and immersive playback approaches be evaluated with respect to the conferencing system?

In order to answer the research questions at the system level of the aspired teleconferencing system, I first have to construct the teleconferencing system with modules that consists of eligible algorithms that are chosen by maximizing local cost functions for the respective modules.

Facing the research efforts in the individual modules of the teleconferencing system, I propose the hypothesis that it is possible to construct a sound acquisition for the teleconferencing system that fulfills the afore stated requirements defined in Section 1.2. Furthermore, I assume that individual or customized HRTF datasets offer a better playback impression of conference contributions compared to a non-individualized HRTF dataset.

Method Description

As illustrated in Figure 1.4, I divided the teleconferencing system into two parts, namely, the sound acquisition and the immersive playback.

The sound acquisition part starts with a review of related work to achieve the channel assignment for the teleconferencing system. Based on the findings, I develop a channel assignment algorithm for the teleconferencing system that consists of a sound source

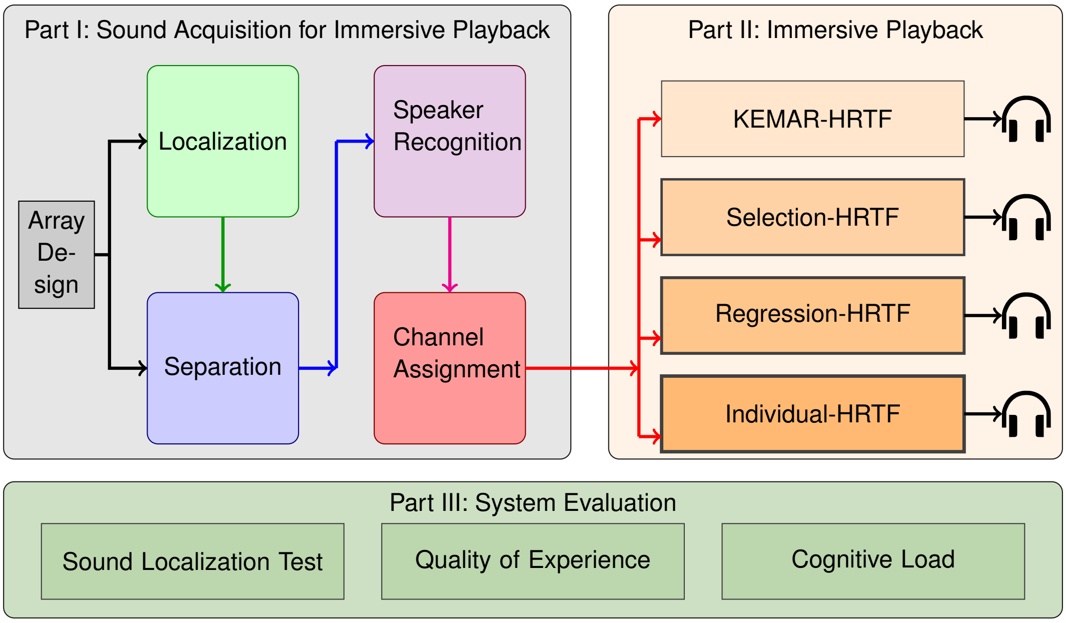

Figure 1.4.: Schematic overview of the modules for the sound acquisition, the immersive playback and the system evaluation

localization, a sound source separation and a speaker recognition module. For each module another literature review is made to choose the most appropriate algorithms for the modules. The sound source localization and separation algorithms are benchmarked by extensive experiments in anechoic and echoic environment with different teleconferencing microphone array prototypes that are constructed with respect to the utilized localization and separation algorithm's properties. The evaluation of the algorithms is done by objective measurements that are frequently utilized in the respective research communities. In order to find parameters such that the speaker recognition system meets the teleconference requirements, extensive simulations with a publicly available meeting corpus are conducted. Finally, the chosen localization, separation and speaker recognition algorithms are combined to a channel assignment system. Different reproducible teleconference situations are designed to have a realistic objective evaluation of the channel assignment system.

In the immersive playback part, I identified three individualization approaches to choose the HRTFs for immersive conference contribution playbacks at the remote site, namely the selection, the regression and the individual measurement approach. For each approach, different algorithms are evaluated by typical measurements of the respective research communities.

Finally, in the evaluation part of my thesis, I identified and conducted three subjective evaluation methods that are appropriate to judge the performance of the teleconferencing system in terms of quality of experience, cognitive load and sound localization accuracy.

1.6. Contributions

The research questions defined in Section 1.5 are answered in this thesis by the following contributions:

Sound acquisition

- I identify and implement eligible state of the art approaches for sound source localization, separation and speaker recognition and adjust the algorithms to the constructed teleconference microphone array prototype.
- I develop a channel assignment algorithm which meet the requirements to be employed in a teleconference system.
- The algorithms of the sound acquisition system are extensively evaluated by objective measures, that are typical for the respective research community.

Immersive playback

- I identify and implement HRTF individualization and acoustic HRTF measurements methods which allow for different degrees of individualization.
- I develop regression-based HRTF customization algorithms using multiway array feature extraction methods.
- I develop a method to fairly compare acoustically measured HRTFs obtained by different measurement approaches.
- The different HRTF individualization methods are evaluated and measurement results, typical for the respective community, were obtained.

Evaluation

- I have planned and constructed a well-equipped semi-anechoic audio laboratory at the Institute for Data Processing (LDV) to evaluate the sound acquisition algorithms in both, anechoic and echoic condition.
- I set up the conditions to acoustically measure individual HRTFs in the audio laboratory at the Institute for Data Processing and succeed to construct the LDV HRTF database that consists of 35 subjects.

- My research team and I identify, develop, conduct and evaluated listening tests to evaluate the listeners sound localization accuracy, the quality of experience and the cognitive load of the developed teleconferencing system.
- The listening tests unveil that the developed channel assignment algorithm works audibly well.
- The listening tests unveil that the test subjects prefer spatial separated playback of the remote conferees.
- The listening tests unveil that individualized or even acoustically measured sets of HRTFs of the probands do not increased the perceived quality of experience of the teleconferencing system.
- The cognitive load listening tests unveil, that HRTF-based playback of conference contributions improve effectiveness and efficiency of a teleconference.

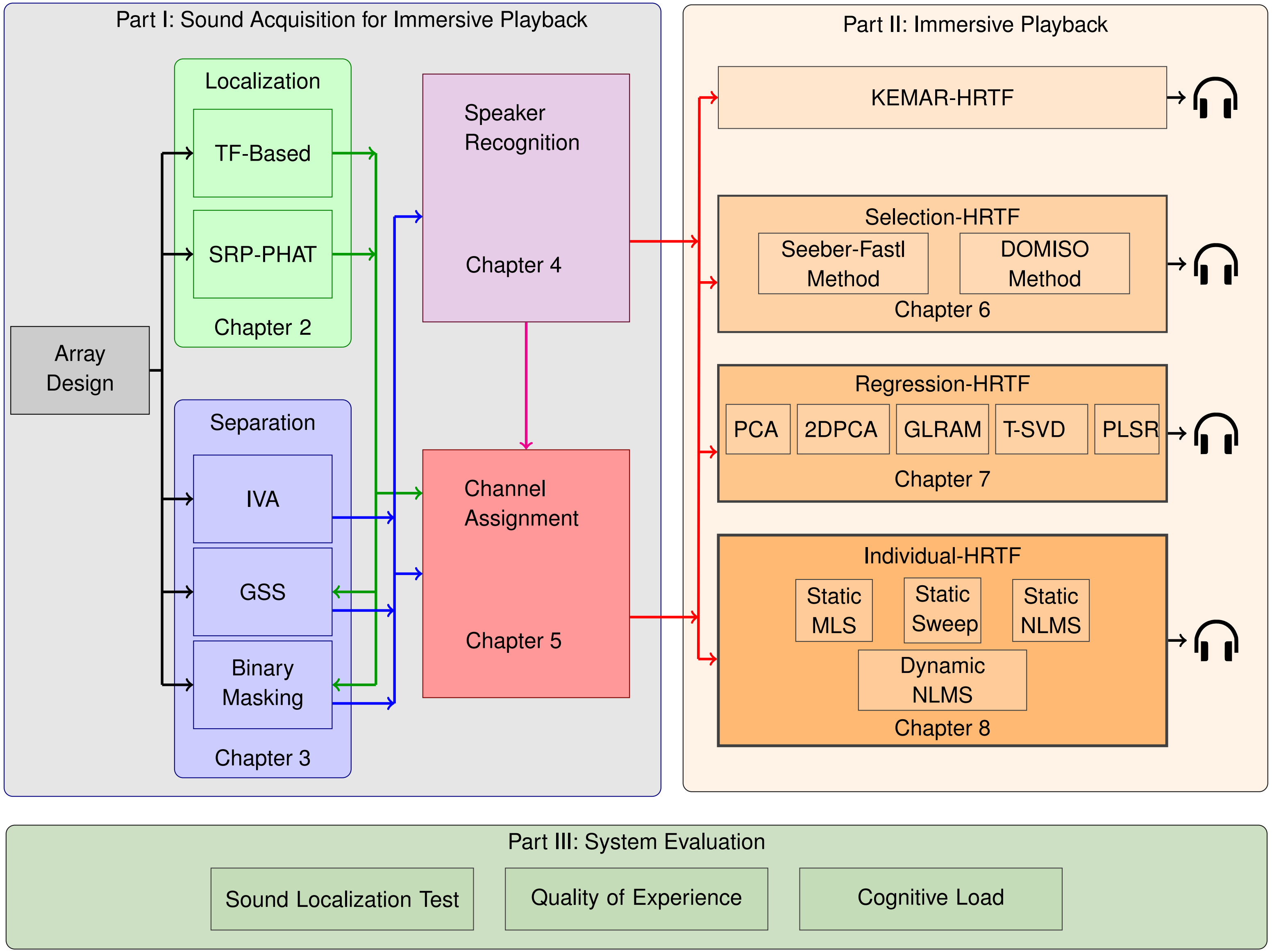

Figure 1.5.: System overview of the modeling parameters of the teleconference system

1.7. Overview

The organization of the thesis follows the signal flow of the aspired teleconferencing system. A schematic overview of the modules of the system is illustrated in Figure 1.5.

Part I of the thesis takes place in the conference room. The speech contributions of the conferees in the conference room are recorded by microphone array prototypes which are designed in Chapter 2 with respect to the evaluated localization algorithms, also presented in Chapter 2. Two different localization algorithms are considered in the thesis, namely the steered response power phase transform (SRP-PHAT) localization algorithm and a transfer function based sound localization approach (TF-based). Two of the three evaluated sound source separation approaches, the geometric source separation (GSS) and binary masking, overviewed in Chapter 3, use the localization information to separate sound mixtures, e.g., in situations where conference participants in the conference room are simultaneously talking. The third separation algorithm, called independent vector analysis (IVA) works without any localization information. The localized and separated signals are then passed to the speaker recognition system presented in Chapter 4 which decides who actually is speaking based on voice features. In Chapter 5, the findings of the sound source localization, the sound source separation and speaker recognition experiments are combined to an algorithm that successfully accomplishes the channel assignment.

Part II takes place at the remote conferee's site. The channel assigned conference contributions of the conference room are virtually synthesized to the remote participant by using HRTF-based sound rendering. The choice of the utilized HRTF dataset for 3D playback can be made on the basis of a costbenefit analysis: The playback without any extra effort by the conferee can be done with a non-individual KEMAR dummy head HRTF. With little extra effort, the listener can achieve an individualization by selecting an appropriate HRTF dataset within an HRTF database. Two related HRTF selection approaches are presented in Chapter 6. A higher degree of individualization can be achieved by applying regression techniques, presented in Chapter 7, to generate an individual set of HRTFs. The regression method is connected to a higher effort than the selection method, since the regression method requires some anthropometric measures. The most individual HRTF, however, can be achieved by acoustically measure the conferees head-related transfer function. In Chapter 8 a method is presented and compared to other approaches to efficiently obtain an acoustically measured set of HRTFs for the remote listener of the teleconference.

Finally, Part III of the thesis gives an overview of three different methods to evaluate the modules and the combination of the modules of the developed teleconference system with respect to the listener and with respect to the task of teleconferencing. Of course, there are already experiments for the respective modules of the system, e.g. determining the diarization error rate for the assignment algorithms, but the objective criteria that were used to judge the single modules and algorithms of the system are in my opinion not suited to evaluate the teleconference system. Therefore, more meaningful evaluation concepts are introduced and applied to the teleconferencing system in Part III of this thesis. One frequently used method to evaluate the performance of a set of HRTFs is the sound source localization ability of a listener. The precise determination of the proband's sound source localization performance itself is a challenging task and requires a complex measurement setup which is developed in the respective section of the thesis. Measuring the sound source localization accuracy of a proband allows for the evaluation of the different playback options, but does not include the sound acquisition of the developed system. Moreover, the pure sound localization accuracy of the played back conference contributions might be not an appropriate measure to quantify the benefit of the developed teleconferencing system. Thus, the quality of experience concept is applied to also evaluate the whole interaction of the different modules for the task of teleconferencing and to identify the playback options that provide the best cost-benefit-ratio for the aspired teleconferencing scenario. This playback options are finally applied for a cognitive load evaluation of the teleconferencing system to determine the effectiveness and efficiency of the developed teleconference system.

This thesis can be regarded as an executive summary of the conferencing system development. In each chapter, I summarize the essential aspects of the respective modules and the achieved results are presented. My research team and I provide more in-depth informations about each chapter in corresponding publications.

Part I.

Sound Acquisition for Immersive Playback

The cocktail-party effect [41] describes the human ability to understand one particular speaker, even if there are simultaneously active speakers. My goal is to develop a system that allows to exploit this ability for teleconferencing. The basis for this system is to develop a sound acquisition system that automatically detects the identity of the active conference contributor and assigns the speech contributions to individual audio channels, which can be virtually synthesized at the remote conference participant site to different positions.

The problem of channel assignment can be straightforwardly solved by equipping each conferee with a close-talking microphone that captures the respective conference contribution [8]. The identification of the conference participants and consequently the assignment is done by the microphone ID that is characteristic for each conferee.

A tabletop microphone array can be regarded as an alternative to the close-talking microphones. However, extra effort has to be spent to assign the conferees to their individual transmission channel, e.g., using sound source localization or speaker recognition algorithms. Sound source localization detects the position of an active conference participant, which can be used to assign the conference contribution on the basis of the localization information at each time instance. The assignment is therefore not conducted by the active conferee's voice, but rather based on the conferee's position. An example for a localization driven diarization of conferences is given in [5]. Furthermore, localization in conjunction with a tracking approach and clustering of the known position of the conference participants [155] can be applied to solve the assignment of the active speaker.

Besides sound source localization, speaker recognition approaches are available for the assignment of the conference participants in the conference room. Compared to sound localization driven approaches exploiting the sound source position for assignment, speaker recognition algorithms seek to identify the conferee by individual voice features.

The problem of channel assignment can also be solved with speaker diarization systems that detect who is speaking when in a conference by using all possible information within the audio recordings, e.g. voice features and direction of arrival of the voice source. Speaker diarization approaches usually work on the recordings of the whole conference to classify and cluster the conference contributions [166] and to reach better assignment results than online speaker recognition algorithms but the offline processing requirement renders diarization systems useless for teleconferencing applications.

Another class of approaches for channel assignment is to combine audio information with camera information, e.g., to monitor meeting participants [26] or to apply reinforcement learning for speaker recognition, as I proposed in [132].

In the first part of my thesis, methods to achieve the channel assignment are presented, eligible algorithms for channel assignment are chosen and adjusted or developed and finally evaluated by real world experiments in a semi anechoic environment and in echoic environments. Furthermore, my research team and I construct hardware prototypes taking the aspired conference scenario and the chosen algorithms' properties into account. Among the afore mentioned classes of approaches for channel assignment, I decided to develop a combination of sound source localization, separation and speaker recognition in order to reach the requirements for my teleconferencing system and to combine the advantages of the respective algorithms for channel assignment in a meaningful way. In Chapter 2 two robotic sound source localization algorithms are introduced and evaluated. The first algorithm utilizes a Steered Response Power - Phase Transform (SRP-PHAT) sound localization and has been proven of value in the field of robotic sound localization using microphone arrays. The second algorithm follows the lead of human hearing and exploits spectral cues to localize sound sources with two microphones. The algorithm is extended to be used with a teleconferencing prototype array that consists of eight microphones. Extensive experiments are conducted with different microphone array prototypes to find an eligible localization algorithm to be used in the channel assignment system. In Chapter 3 three different sound source separation algorithms are presented and evaluated regarding the task of teleconference channel assignment. The first algorithm is a blind sound source separation algorithm, called Independent Vector Analysis (IVA) which separates the sound mixtures by statistical models without any localization information of the sound source. The second algorithm is named Geometric Source Separation (GSS) and additionally uses localization data to improve the sound source separation. The third algorithm is called binary masking and separates the mixtures solely by localization data. Chapter 4 introduces a speaker recognition system and the parameters of this system for the task of teleconferencing are determined by numerous experiments. Finally, an assignment system is constructed and evaluated in Chapter 5 with the chosen sound source localization, sound source separation and speaker recognition algorithms.

2. Sound Source Localization

Sound source localization (SSL) algorithms using microphone arrays are the standard approaches to detect sound sources in various fields of application, e.g., robotic sound source localization. According to [22] microphone array based sound localization algorithms can be classified into three groups, namely, time delay of arrival (TDOA) based approaches, high-resolution subspace techniques (HRST) and steered response power (SRP) beamforming algorithms.

The TDOA based SSL approaches localize a sound source by exploiting the time differences of the impinging sound waves between the array's microphones [37, 150]. By knowing the microphone coordinates, one can estimate areas of possible sound source locations, which are refined by an increasing number of different microphone pair's time delays. Usually, the generalized cross-correlation method [90] is applied to compute the TDOAs between the pairs of microphones. Regarding echoic environments or scenarios with more than one simultaneously active sound source in an echoic environment, evaluating the TDOAs of the cross-correlation estimation is a challenging task [149].

HRST methods are also named spectral-estimation-based locators and make use of the so-called cross-sensor covariance matrix. The cross-sensor covariance matrix is a correlation matrix computed across the spatially distributed microphones and is used to estimate parameters that influence the microphone recordings. One popular HRST method is the multiple signal classification (MUSIC) method [151]. The basic idea of MUSIC is the assumption that the recording of each microphone is a linear combination of a sound source and disturbing noise. A principal component analysis on the covariance matrix computes the disjoint signal subspace and noise subspace. MUSIC also estimates the most likely sound source directions by omnidirectional distance evaluation of the array's steering vectors and the noise subspace. Another example for HRST methods is ESPRIT [146]. The restrictions for applying HRST methods in a teleconference system are the computationally expensive calculations for the principle component analysis and the requirement of a higher number of microphones than the number of sound sources including echoes and noise sources.

Beamforming is a frequently used microphone array sound localization technique to add a directivity to the microphone array. Therefore, the array that consists of omnidirectional microphones can intensify sound signals from a certain direction by so-called constructive interferences and suppresses noise and echoes from other directions by destructive interferences. The interferences are generated by processing each microphone's recording by time shifts according to the physical microphone array setup and according to the steering angle. The time shifted versions of the recordings are then summed up to an output signal and constitute constructive and destructive interferences. This technique is denoted as delay-and-sum beamformer [22]. Additional frequency dependent weighting is applied in a filter-and-sum beamformer to the time shifted microphone recordings. A SRP beamformer localizes a sound source by steering the directivity such that the highest output power of the beamformer is reached among predefined search positions around the microphone array [21].

The transfer function based approach can be regarded as an alternative approach to microphone arrays. The idea is to exploit the human cues for SSL, which can be adapted for the teleconferencing system. According to human-like HRTF-based sound localization, sound waves approaching the microphone array are diffracted and reflected of the array's shape such that direction dependent spectral changes can be observed. Furthermore, microphones also offer direction dependent transfer behavior denoted by acoustic transfer functions (ATF), which can be roughly compared to HRTFs. Teleconference systems could benefit from a human-like sound localization approach because of the ability to localize sound sources in a three-dimensional environment with only two microphones in a compact manner. Recently, ATF-based sound localization algorithms have been developed to enable mobile robotic platforms to localize sound sources. In [44, 86, 104] sound source localization algorithms based on HRTFs have proved to be accurate and robust to noise.

2.1. Teleconference Sound Localization

In my opinion the SRP sound localization method or the ATF-based sound source localization method are beneficial to be employed in a teleconferencing system since the algorithms have proved valuable in robotic sound source localization applications. After briefly summarizing the algorithms, extensive experiments are designed, conducted and evaluated to choose the most promising approach for the final teleconferencing system.

Steered Response Power (SRP) Localization

This section describes our adaption of localizing and separating sound sources, based on [170, 172].

The SRP-PHAT algorithm [39] can be regarded as one promising approach to localize one or multiple simultaneously active sound sources in an echoic and noisy environment. The SRP-PHAT algorithm is therefore also utilized for robotic sound source localization in [170, 172]. SRP-PHAT estimates the TDOA by determining and aligning the cross-correlation functions of the microphone signals of the array according to pre-computed delays and the PHAT weighting function. The SRP P impinging from a certain azimuth angle φ and elevation angle θ of the filter-sum beamformer is computed by

$$P_{\varphi,\theta} = \sum_{l=1}^{N} \sum_{k=1}^{N} \int_{-\infty}^{+\infty} \Psi_{lk}(f) x_l(f) x_k^*(f) e^{j2\pi f(\tau_k - \tau_l)} df, \tag{2.1}$$

where

$$\Psi_{lk}(f) = \frac{1}{|x_l(f)x_k^*(f)|} \tag{2.2}$$

is the PHAT weighting and τ_k and τ_l are the (φ, θ)-dependent delays between a pair of microphone signals $x_l(f)$, $x_k(f)$ at a sampling point of the search region [39]. The search region of the localizer can be described as a grid of points that covers preselected possible directions of arrival. $P_{\varphi,\theta}$ has to be computed between each microphone pair and sampling point. The summation of the possible $P_{\varphi,\theta}$ results in an energy map that has peaks at those positions, where sound sources are expected. Since we focus on teleconferencing applications with the microphone array placed on a conference table, we restrict the search region to the upper hemisphere.

Transfer Function Based Sound Localization

The transfer function (TF) based sound source localization algorithms returns the azimuth angle φ and the elevation angle θ of the sound source using the recorded microphone signals and a stored database of the microphone array's acoustic transfer functions (ATFs). The unknown signal s emitted from a sound source is convolved by the corresponding ATFs denoted by a_{i,φ_0,θ_0} before being captured by the microphones of the microphone array, i.e.,

$$x_i = a_{i,\varphi_0,\theta_0} * s, \tag{2.3}$$

where φ_0 and θ_0 denote the azimuth and elevation angle corresponding to the direction of the sound source s and $*$ is the convolution operator.

One technique is based on the fact that filtering x_i with the inverse of the correct ATFs $a^{-1}_{i,\varphi_0,\theta_0}$ yields identical signals $\widetilde{s}_{i,\varphi,\theta}$.

In order to avoid inversion caused instabilities [86, 104], the cross-convolution approach exploits the associative property of the convolution operator [104, 139, 169] by

$$\widehat{s}_{1,\varphi,\theta} = a_{2,\varphi,\theta} * x_1 \tag{2.4}$$

and

$$\widehat{s}_{2,\varphi,\theta} = a_{1,\varphi,\theta} * x_2, \tag{2.5}$$

which turn to be identical at the correct source position for the ideal case, i.e.,

$$\widehat{s}_{1,\varphi,\theta} = \widehat{s}_{2,\varphi,\theta} \Longleftrightarrow (\varphi, \theta) = (\varphi_0, \theta_0). \tag{2.6}$$

The source can be localized in real applications by

$$\underset{\varphi,\theta}{\operatorname{argmax}} \left\{\widehat{s}_{2,\varphi,\theta} \oplus \widehat{s}_{1,\varphi,\theta}\right\}, \tag{2.7}$$

where $\oplus$ denotes a cross-correlation operation. In [44] the cross convolution based sound localization approach using measured transfer functions for a pair of microphones is extended to multiple active sound sources. The approach exploits disjoint sets of Fourier transform supports since the W-disjoint orthogonality [128] of different simultaneously active sound sources, that are assumed to be sparse in some transform domain, can be utilized. Therefore, only one active source at each time-frequency bin can be observed at each time-frequency point after a Fourier transform. Comparable to the cross convolution algorithm, a similarity is computed for each frequency bin. The most likely transfer function indices for each frequency bin are stored and weighted resulting in a histogram, where the peaks correspond to the position of the active sound sources.

The TF-based algorithm works basically on the pairwise comparison of two microphone's recordings. To exploit redundant recording information of all eligible microphones of the teleconferencing array, subgroups of two microphones out of the eight microphones of the array can be utilized for SSL. Finally, histograms of each microphone pair are fused to one compound histogram.

Particle Filtering

The stand alone SRP-PHAT localizer and the TF-based approach produce instable sound localization results, including localization of noise sources and echoes. To overcome this problem, a particle filter is integrated [73] for temporal smoothing of noisy measurements. Every possible sound source is therefore considered to be a set of particles, where each particle is assigned to a distinct position in space, velocity and weighting. The sound localization estimations of the localization algorithms are then used to update the particle positions, directions and weightings, resulting in a permanently updated probability density function (PDF) of the estimated sound sources. By computing the mean value of the PDF, a stable sound source localization can be achieved.

2.2. Design of the Array

Several information cues, such as time delay of arrival (TDOA), need to be considered in designing the physical setting of the arrays [21]. In microphone array-based sound localization approaches, the size of microphone arrays varies drastically. Some designs are compact [99, 171], others yet are cumbersomely large. Some of the investigated microphone arrays have several meters in diameter [164] and would be too big for our aspired scenario. It is also vital to carefully determine the bandwidth requirements to design the hardware geometry [40, 47]. Choosing a microphone array geometry depends primarily on the intended spatial coverage. In literature, there exist various frequently used geometries of arrays, such as linear, square [171], circular [163] and spherical [47] arrays. Some applications, such as a teleconference system, that is placed on a table, may only need to define the direction in the upper hemisphere, which allow for a simpler geometry than an microphone array requiring full unambiguous determination of a sound source's azimuth and elevation. The number of used microphones scales spatial resolution and noise robustness. Likewise, arrays that are equipped with a huge number of microphones may result in high accuracy of localization, however, they require extensive calibration and high-speed dedicated multi-channel hardware capable of handling high data throughput.

The distance between the array's microphones also is an issue due to the aliasing problem. To avoid spatial aliasing, the distance between the microphones should be smaller than half of the minimum wavelength of the incoming sound signal. In [50] we overview possible array configurations for our teleconferencing system, e.g., harmonically nested subarrays, linear arrays, planar arrays and volumetric arrays. With regard to our teleconference scenario, a circular microphone array design consisting of eight microphones was selected. The localization accuracy experiments that we did, revealed that a planar geometry has no significant disadvantages in our scenario compared to a volumetric circular array configuration. Based on the wooden circular microphone prototype in [50], my research team and me designed and constructed three plastic microphone arrays, each of them consists of eight low budget microphones [58].

One planar array, illustrated in Figure 2.1, is constructed to fit the SRP beamformer. The microphones have a direct line of sight to each other, therefore, time delays can directly be estimated by knowing the microphone positions.

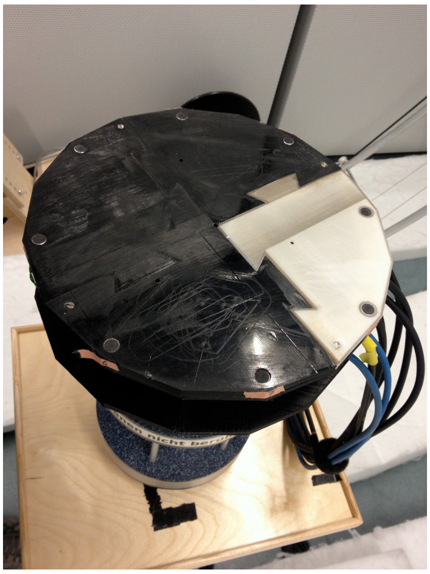

Figure 2.1.: Microphone Array 1: The planar shape of the array is constructed with respect to the SRP-PHAT algorithm

Figure 2.2 shows the second prototype which is constructed to fit the TF-based approach. The TF-based approach utilizes spectral direction dependent differences to localize sound sources. So far, a human-like head, shoulder, torso and pinna shape was required in order to use the TF-based approach [43]. To meet the algorithm's requirement of direction dependent features in our teleconference scenario, we construct a microphone array with a conchiform sound reflector for each microphone of the array. The concha is constructed as suggested in [100]. By a cutout of an Archimedean spiral, the reflector behind the microphone has different distances of the microphone for every angle, consequently the reflections caused by the concha are direction dependent. The resulting direction dependent peaks and notches in the spectrum are beneficial for the transfer function based localization algorithm. Finally, a third array, illustrated in Figure 2.3, is built that introduces a level difference between the microphone recordings, which might be useful for the transfer function based approach and for the sound separation process. The design of the third microphone is less obtrusive than the concha design of the second array.

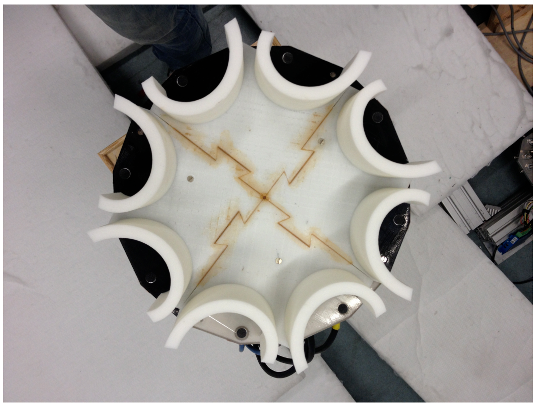

Figure 2.2.: Microphone Array 2: The array shape is constructed with regard to the TF-based localization approach

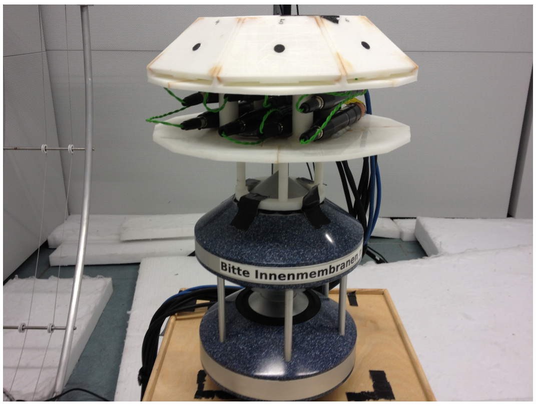

Figure 2.3.: Microphone Array 3: The array shape reinforces a level difference between the microphone recordings

Sound sources	KS digital C5 tiny
Microphones	CUI CMB-6544PF
Microphone preamplifiers	Focusrite Sapphire Pro 40
Sound card	RME Multiface II
Anechoic room dimensions	4.7 m x 3.7 m x 2.84 m
Echoic room dimensions	6.3 m x 4.0 m x 2.8 m
Anechoic room noise level (A-weighted)	16.3 dB
Echoic room noise level (A-weighted)	20.9 dB
Anechoic room reverberation time t_{60}	0.08 s
Echoic room reverberation time t_{60}	0.23 s

Table 2.1.: Information about the equipment and the environment used in our localization experiments.

2.3. Sound Source Localization: Experiment

In this section, we apply the SRP-PHAT and the TF-based sound source localization algorithm to a conference situation based sound localization problem. The localization algorithms are extensively evaluated by real world recordings in a semi anechoic chamber and by real world recordings in an echoic environment.

Sound Source Localization: Experimental Settings

Two different scenarios for the experiments are considered for the three microphone array prototypes. In the first scenario, the sound source localization algorithms are tested by conducting a conference in the semi anechoic audiolab. Then the same experiments are held in an echoic environment.

To guarantee a fair comparison among different algorithms and settings, sound scenes are prerecorded. Therefore, instead of human conference participants, loudspeakers are placed in a distance of 1.3 m at three different angles (45°, 135° and 225°) around the conference table. The microphone array is placed in the center of the table. Speech contributions (10 s) of eight male and four female speakers were recorded and played back in different settings with the loudspeakers. In sum, 1008 different recording situations are covered for the anechoic and echoic scenario. The room characteristics and information about the equipment used in the localization experiments are given in Table 2.1.

Sound Source Localization: Experimental Results

In each experiment, the localization performance of the two sound source localization algorithms is evaluated for the three teleconferencing system prototypes by the mean angular error (MAE) and the localization success rate within a tolerance region of 5° (TOL). The localization success rate is computed by the ratio between the number of localization results within the tolerance region and the total number of frames.

Transfer Function Based Sound Localization (TF)

As we know, the TF-based approach requires measured transfer functions of each possible location of a sound source. Therefore, we feed the localization algorithm with premeasured transfer functions of the anechoic room and utilize the database of transfer functions for the anechoic and echoic conference recordings. Tables 2.2, 2.3 and 2.4 give an overview about the localization performance of the TF-based sound localization approach.

For Array 1 the TF-based localization algorithm achieves good localization performance. The localization accuracy in azimuth at elevation 10° for one active sound source is TOL_{az} = 98.2% with a MAE_{az} = 0.7° in the anechoic environment. Adding a second active sound source, the localization accuracy and the *MAE* for both sources is slightly lower. In the elevation 20° plane, the TF-based approach suffers higher inaccuracies, revealing that the direction-dependent spectral differences for the elevation 20° plane are lower than in the elevation 10° plane of Array 1.

Array 2 is constructed to add direction-dependent spectral differences in the audio recordings of the array. In the elevation 10° plane the MAE_{el} values are better than for the Array 1 recordings as seen in Table 2.3. In the elevation 20° plane the localization performance in both, elevation and azimuth is very good, denoting that the concha-applications improve the performance of the TF-based approach, especially for the elevation 20° plane. Array 3 does not have any further advantages for the TF-based localization approach, which can be seen in Table 2.4.

It is worth mentioning that the localization performance of the TF-based approach in the echoic environment is still remarkable, regarding the fact that the acoustic transfer functions are measured in the anechoic environment.

Steered Response Power (SRP) Localization

Table 2.2 gives an overview of the localization performance of the SRP-PHAT sound localization approach for Array 1. The localization accuracy in azimuth at elevation 10° for one active sound source is TOL_{az} = 99.3% with a MAE_{az} = 0.22° in the anechoic environment. Regarding more than one simultaneously active sound source and echoic environment recordings, the localization performance of the SRP-PHAT sound localization algorithm is still excellent for Array 1.

The concha-applications of Array 2 influence the localization performance of the SRP-based sound localization algorithm to a huge extent, as illustrated in Table 2.3. For example, the MAE_{az} for the elevation 10° in the anechoic room deteriorates to MAE_{az} = 4.2° compared to a MAE_{az} = 0.22° for Array 1. The reason for the performance decrease may be the direction dependent path differences between the microphone introduced by the concha application which cause time delays of impinging sound waves that are not considered in the SRP algorithm that localizes sound sources by exploitation of the time delays assumed by the microphone position. As seen in Table 2.4, Array 3 introduces elevation localization inaccuracies due to the shadowing effects of the construction, while azimuth localization performance is not harmed by the introduced level difference between microphone recordings of Array 3.

2.4. Concluding Remarks: Sound Source Localization

Regarding the audio conferencing scenario, the SRP-PHAT algorithm has the advantage of working without pre-measured acoustic transfer functions and only requires the position data of the microphones. However, the TF-based localization approach performs well in echoic environment with pre-measured acoustic transfer functions of the anechoic environment, which would be an acceptable compromise for the real-world application of the teleconference microphone array.

The choice of the sound localization algorithm and the microphone array shape that is further used in our prototype is consequently made on the basis of the localization performance. With respect to the sound acquisition scenario, I regard the localization performance in azimuth as a proper criterion, because the azimuth position of a conference participant is a more critical factor for the channel assignment and the following immersive playback than the elevation position. The main difference between conference participants, besides voice features, is their location in azimuth. Therefore, the TOL_{az} is a fair evaluation criterion to choose the sound localization algorithm for the conferencing system. In terms of TOL_{az}, the SRP-PHAT approach outperforms the TF-based approach in the majority of the localization experiments. Hence, I decide to use the SRP-PHAT sound localization approach in combination with Array 1 which provides the best localization performance for the SRP-PHAT approach and furthermore, is the most unobtrusive construction among the array shapes. My research team and I provide more information about the localization algorithms and further experiments in [50, 58, 134, 138].

Algorithm	Anechoic				Echoic			
	MAE_{az}	MAE_{el}	TOL_{az}	TOL_{el}	MAE_{az}	MAE_{el}	TOL_{az}	TOL_{el}
Elevation: 10°								
one source								
TF-based	0.7°	1.4°	98.2%	91.4%	1.3°	6.6°	96.6%	32.2%
SRP-PHAT	0.22°	4.3°	**99.3%**	84.9%	1.0°	4.6°	**97.6%**	74.9%
two sources								
TF-based	1.8°	2.8°	**95.6%**	74.5%	6.8°	7.7°	90.8%	21.7%
SRP-PHAT	0.3°	3.8°	95.2%	82.3%	0.5°	4.1°	**93.5%**	74.5%
Elevation: 20°								
one source								
TF-based	1.1°	1.5°	96.2%	89.6%	1.7°	6.9°	94.1%	44.6%
SRP-PHAT	0.8°	2.5°	**97.1%**	96.5%	0.8°	2.3°	**98.1%**	97.1%
two sources								
TF-based	11.5°	3.0°	87.3%	77.9%	10.7°	10.9°	88.0%	30.9%
SRP-PHAT	0.6°	2.5°	**94.4%**	90.3%	0.6°	2.3°	**92.7%**	89.3%

Table 2.2.: Array 1: Comparison of the different localization approaches with recordings from the planar array [58]. TOL_{az} and TOL_{el} describe the elevation and azimuth localization success rate within a tolerance region of 5°. MAE_{az} and MAE_{el} are the respective mean angular errors.

Algorithm	Anechoic				Echoic			
	MAE_{az}	MAE_{el}	TOL_{az}	TOL_{el}	MAE_{az}	MAE_{el}	TOL_{az}	TOL_{el}
Elevation: 10°								
one source								
TF-based	2.2°	0.9°	**87.0%**	95.7%	11.7°	4.3°	**77.9%**	59.6%
SRP-PHAT	4.2°	12.1°	65.4%	12.9%	6.4°	9.7°	47.1%	20.9%
two sources								
TF-based	7.3°	1.9°	**74.1%**	83.7%	50.7°	5.2°	**46.4%**	52.4%
SRP-PHAT	6.0°	10.0°	39.8%	18.7%	6.0°	9.1°	40.9%	21.3%
Elevation: 20°								
one source								
TF-based	0.7°	0.6°	**97.9%**	97.9%	1.1°	0.7°	**97.3%**	97.9%
SRP-PHAT	3.9°	20.2°	85.2%	0.39%	4.3°	19.1°	74.3%	2.5%
two sources								
TF-based	0.8°	0.7°	**95.6%**	95.6%	7.0°	1.3°	**88.4%**	92.6%
SRP-PHAT	4.0°	18.6°	63.2%	2.8%	5.1°	17.2°	45.5%	4.5%

Table 2.3.: Array 2: Comparison of the different localization approaches with recordings from the concha-microphone array [58]. TOL_{az} and TOL_{el} describe the elevation and azimuth localization success rate within a tolerance region of 5°. MAE_{az} and MAE_{el} are the respective mean angular errors.

Algorithm	Anechoic				Echoic			
	MAE_{az}	MAE_{el}	TOL_{az}	TOL_{el}	MAE_{az}	MAE_{el}	TOL_{az}	TOL_{el}
Elevation: 10°								
one source								
TF-based	2.1°	3.8°	87.9%	64.5%	9.6°	8.6°	85.0%	12.5%
SRP-PHAT	0.2°	5.6°	**99.3%**	44.8%	1.8°	7.2°	**96.4%**	15.0%
two sources								
TF-based	13.6°	3.6°	73.3%	66.9%	16.5°	8.8°	77.9%	11.4%
SRP-PHAT	0.6°	6.4°	**95.2%**	31.2%	0.8°	7.0°	**91.9%**	14.6%
Elevation: 20°								
one source								
TF-based	1.1°	1.7°	94.9%	88.1%	9.4°	7.9°	85.5%	41.6%
SRP-PHAT	2.1°	16.3°	**96.9%**	0.47%	1.6°	16.2°	**97.7%**	0.57%
two sources								
TF-based	4.7°	3.0°	94.2%	82.1%	9.7°	10.2°	83.6%	39.5%
SRP-PHAT	1.0°	17.3°	**95.4%**	0.1%	1.2°	17.2°	**93.8%**	0.1%

Table 2.4.: Array 3: Comparison of the different localization approaches with recordings from the shield-microphone array [58]. TOL_{az} and TOL_{el} describe the elevation and azimuth localization success rate within a tolerance region of 5°. MAE_{az} and MAE_{el} are the respective mean angular errors.

3. Sound Source Separation

To enable a technical system to focus on one specific sound source within a mixture, source separation techniques are required that process the observed mixture into its underlying signal parts. The term blind source separation (BSS) that is often mentioned in sound source separation scenarios, refers to methods for the estimation of source signals using only information acquired by the analysis of recorded mixtures. This excludes prior information, for example, about the frequency characteristics, the location or the mixing process. Yet some information like the location of the sound source can be obtained by sound source localization and be used to improve the performance of separation algorithms. Up to now, numerous algorithms for the separation problem have continuously evolved. These algorithms can be divided into two groups: The first group uses spatial information, the second group statistical information of the signals to achieve separation.

Beamforming for example applies a spatial filter to separate signals which originate from different locations by linearly combining spatially sampled time series of the sensor data [173]. Beamforming can be combined with direction of arrival (DOA) estimation algorithms like MUSIC [151] or ESPRIT [146] to attain segregated signals. Another algorithm that exploits spatial information is applying spatial spectral masking. Spatial spectral masking first finds the DOA for different sources for each frequency. Afterwards, a spatial filter in the frequency domain is applied [102].

Besides the algorithms that use spatial information, there are methods based on the evaluation of the signals' statistics. Independent component analysis (ICA) [36, 71] is one of the most popular approach of this group of algorithms that successfully perform the BSS for instantaneous mixtures. However, ICA cannot separate convolutive mixtures. To overcome this problem, statistical separation methods that are based on ICA extend its capabilities to tackle convolutive mixtures. This is usually done by transforming the convolutive mixture into the frequency domain, which results in an instantaneous mixture model per frequency bin. Due to the inherent ambiguity of ICA, namely permutations and scalings, frequency domain based ICA ends up with misalignment between frequency bins.

To sort out these misalignments, algorithms like multidimensional independent component analysis (MICA) [30] or independent subspace analysis (ISA) [32] seek to group dependent scalar mixtures and thus achieve the desired separated signals.

When using these algorithms for speech separation, several problems arise. First, the mixing model of ISA and MICA is not designed to fit realistic mixing conditions for speech signals in reverberant environments. For speech mixtures, the assumption holds that only signal parts within the same frequency interval are mixed due to the signal propagation in real environments which usually do not alter the frequency of certain signal parts. Therefore, the mixing model of MICA/ISA does not perfectly fit for speech separation, as it allows for both arbitrary mixing of frequencies and different numbers of scalar variables within the outcomes. MICA/ISA is actually designed to have one large mixing layer for all frequencies, i.e. the highest frequencies are allowed to mix with the lowest frequencies. Second, due to these extensive mixture models, MICA and ISA are complex and computationally expensive algorithms.

Recently, a promising approach called independent vector analysis (IVA) has been proposed to inherently solve the permutation problem [94]. Although the basic ideas behind IVA resemble MICA and ISA, its mixing model is designed especially for the task of audio source separation, grouping dependent frequencies of sources together within the separation step. Also, IVA is not as computationally expensive as MICA or ISA, as the mixing model is simpler allowing for fast computation of the outcomes.

If confronted with a situation where there are more sources than microphones, i.e., the underdetermined case, the general model of ICA and IVA needs adjustments, refined assumptions of the underlying model, or preprocessing to work [175]. Popular algorithms for the underdetermined case assume W-disjoint orthogonality, meaning that sources do not or do rarely overlap in the time-frequency domain. With this assumption, methods like DUET and its expansions achieve underdetermined source separation [110]. DUET-based algorithms first filter out sources by their DOA until a determined problem is reached and then estimate a demixing matrix by the DOA to separate the signals. The requirement of W-disjoint orthogonality is relaxed if a method like TIFROM is used where only small time and frequency segments are used for the estimation of the demixing matrix, whereas DUET operates over the complete time-frequency plane [1].

3.1. Teleconference Sound Source Separation

For the teleconference scenario with a tabletop microphone array that consists of eight microphones, we assume that we have to deal with the overdetermined sound source separation since the array's microphones outnumber the simultaneously active conference participants in one conference room.

In the following, three representatives of different sound source separation techniques are tested for their appropriateness in the conferencing prototype. The first candidate is the afore mentioned IVA, a representative of the separation algorithms that use statistical information without any knowledge of the location of the sound sources.

Beside statistical information, spatial information of the sound sources, provided by the SRP-PHAT localization, can be exploited to separate the mixture. The second algorithm is based on the assumption of sparsity of the signal spectra and uses localization information to mask the individual sound sources out of the mixture. Therefore the algorithm is called binary masking. The third observed algorithm, namely geometric source separation (GSS), seeks to combine benefits of BSS and available location information.

Sound Mixing Model

For a scenario of N active sound sources $s_1(t), \ldots, s_N(t)$ and j microphones that capture the mixtures $x_1(t), \ldots, x_j(t)$, the most intuitive mixture model is the instantaneous mixture model,

$$x_j(t) = \sum_{i=1}^{N} a_{ij} \cdot s_i(t), \tag{3.1}$$

where, a_{ij} describes an attenuation factor due to different volumes of the sources at each microphone. This instantaneous mixture model is one of the basic models used for independent component analysis (ICA). For these instantaneous mixtures, a huge variety of separation algorithms exist, for example FastICA. An overview can be found in [71] and [31].

For audio signals however, there are better mixing models that take physical properties of sound and the environment into account. Besides the afore mentioned volume differences, it is advantageous to also exploit time lags within the recordings, which occur between the microphones. Furthermore, reverberant rooms often render instantaneous mixture models useless because of time-delayed and scaled versions (echoes) in the microphone recordings. Scaling, time-delay and echoes can be altogether described as a linear filter which is applied to the sound source. Applying a filter mathematically means convolving the original sound source with the corresponding impulse response that is dependent on the position of the sound source and the microphone within a reverberant room. According to the previous instantaneous mixing model, the filter functions are denoted by $a_{ij}(t)$. For microphone j, the recorded signal is the superposition of the filtered sources, computed by

$$x_j(t) = \sum_{i=1}^{N} a_{ij}(t) * s_i(t), \tag{3.2}$$

where $*$ denotes the convolution operation. In literature, this model is called the convolutive mixture model. The convolutive mixture model allows for a better description of the mixing process for sound sources and consequently enables us to reach better sound source separation results.

There are a number of algorithms that perform separation on the mixtures directly [106] which are computationally expensive. To circumvent this problem, the convolutive mixture model can be transformed to frequency domain, where the convolution operation is described by a multiplication and consequently a convolutive mixture model (3.2) turns into an instantaneous mixture model

$$x_j(f) = a_{1j}(f) \cdot s_1(f) + \cdots + a_{Nj}(f) \cdot s_N(f), \tag{3.3}$$

which is similar to Equation (3.1). In contrast to the time domain instantaneous ICA model, the mixing coefficients are also dependent on the frequency variable, which renders the direct use of instantaneous mixture algorithms such as FastICA useless. The frequency domain representation of the model can be described by

$$x(f) = A(f) \cdot s(f). \tag{3.4}$$

It is obvious that Equation (3.4) indeed corresponds to the instantaneous mixture model with the flaw that there is not one mixing matrix A, but one for each frequency bin $A(f)$.

Independent Vector Analysis (IVA)

In contrast to solve the permutation problem after demixing, IVA seeks to avoid the permutation problem within the separation process itself by the assumptions that the components of a source over all frequency bins are dependent and the different source's components within one frequency bin are independent. Therefore, IVA is capable of solving the permutation problem inherently [94]. Taking this into account with a suitable cost function, the IVA algorithm manages to identify the dependent frequency components of each source.

Prior to the separation process, the frequency bin mixtures are whitened, i.e. they are transformed to uncorrelated mixtures and the bins are assigned to the same variance (power) [71].

In order to apply IVA, certain source priors, called probability distributions $p_i(s_i)$ have to be assumed. The spherically symmetric Laplacian distribution is proposed as source prior. It was shown in [95] that the distribution allows for good approximation of speech and is capable of modeling dependencies among frequencies.

The source priors are then utilized to construct a likelihood-maximizing cost function. It is assumed that the whitened mixtures $x_0(f)$ are separated by a demixing matrix $W(f)$ to yield the estimates $\hat{y}(f)$ for each frequency bin, computed by

$$\hat{y}(f) = W(f)x_0(f) \tag{3.5}$$

and the estimates $\hat{y}(f)$ are combined to $\hat{y}_i$.

IVA seeks to compute a set of demixing matrices $W(1)....W(F)$ that separate the mixtures according to the distribution of the source prior. By maximizing the likelihood L of the estimates $\hat{y}_i$ of the source distribution, i.e.,

$$\underset{W(1),...,W(F)}{\operatorname{argmax}} \; L(W(1),...,W(F)) = \underset{W(1),...,W(F)}{\operatorname{argmax}} \sum_{i=1}^{N} \ln(p(\hat{y}_i)) \qquad \text{s.t. } W(f)W^H(f) = I \quad \forall f \tag{3.6}$$

a function that *measures* the "quality" (in terms of likelihood) of the separation matrices is derived, which is only dependent on the observed mixtures and the separation matrices. Finally, a spectral compensation according to [62] reverses the whitening process and a transformation of the separated mixtures from frequency domain to time domain concludes the blind source separation via IVA. For a more detailed description of applying IVA for robotic and conferencing sound source separation, please refer to our work in [133, 168].

Geometric Source Separation (GSS)

GSS [119] seeks to combine benefits of blind source separation and beamforming by fusing cross-power minimization of convolutive mixtures with geometric information provided by sound localization. In accordance with [119, 170], the cross-talk is minimized by cost functions, given by

$$J_1(W(f) = \|R_{yy}(t,\tau) - diag(R_{yy}(t,\tau)\|^2 \tag{3.7}$$

and

$$J_2(W(f)) = \|W(f)A(f) - I\|^2, \tag{3.8}$$

where $J_1(W(f))$ expresses the cross-talk minimization of the output signals $y(t)$ and $J_2(W(f))$ is the geometric constraint containing the estimated linear transfer functions $A(f)$ between the sources and the microphones. The entries of $A(f)$ are determined by using the sound localization information of the localizer. With $J_1(W(f))$ and $J_2(W(f))$ the separation matrix $W^n(f)$ is updated by

$$W^{n+1}(f) = W^n(f) - \mu \left[\alpha(f)\frac{\delta J_1(W(f))}{\delta W^*(f))} + \frac{\delta J_2(W(f))}{\delta W^*(f))}\right], \tag{3.9}$$

where μ is the adaptation rate and $\alpha(f) = \|R_{xx}(t,\tau)\|^{-2}$ is an energy normalization factor. Finally, the separated output $y(f)$ can be computed by $y(f) = W(f)x(f)$, where $x(f)$ describes the microphone input signals, i.e. the mixtures.

Binary Masking

Beside independent vector analysis that seeks to separate sound mixtures by statistical assumptions and geometric source separation that additionally includes location information of the sound sources, binary masking, as used in [43], seeks to separate signals solely on the location information and the assumption that human speech is sparse in Fourier domain. Furthermore, it is supposed that different speech sources dominate different frequency bins. Therefore, binary masking seeks to separate dominant sound sources at the respective dominant frequency bins.

A binary mask M filters the frequencies attributed to an azimuth angle $\theta(f,m)$ and its circumjacent values and forces the other frequencies to zero for a time point m by

$$M(f,m) = \begin{cases} 1, & \forall\, \theta(f,m) \ \mid\ \theta_{min} \leq \theta(f,m) \leq \theta_{max} \\ 0, & \text{otherwise.} \end{cases} \tag{3.10}$$

For each localized sound source, the binary mask can be applied and the separated signal $\hat{y}$ can be retrieved from a mixture x by

$$\hat{y}(f,m) = M(f,m)\ x(f,m). \tag{3.11}$$

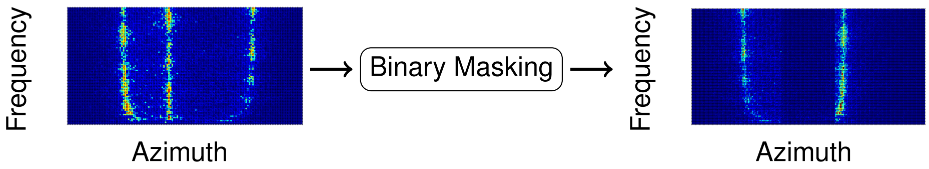

Figure 3.1.: Schematic illustration of the sound source separation using binary masking

Figure 3.1 schematically illustrates the separation of one speech source. After localization a binary mask separates the speech source at a certain azimuth region. One advantage of binary masking is the ability to work also for the underdetermined case and the low computational complexity.

3.2. Sound Source Separation: Experiment

According to the sound source localization algorithms, the sound source separation algorithms, namely, IVA, GSS, and binary masking, are applied to a conference situation motivated sound source separation problem.

Sound Source Separation: Experimental Settings

In accordance to the sound source localization experiments, two experiments are conducted to compare the separation performance. The experimental setting is identical to the sound source localization experiments of Section 2.3. In the first experiment, the two or three simultaneously active loudspeakers play back the conference contribution in the anechoic environment [145]. In the second experiment, the same procedure is repeated in an echoic environment.

Recordings of eight male and four female human speakers are used for the experiment. In sum, 432 sound mixtures are evaluated in the anechoic environment and in the echoic environment, respectively. Contrary to binary masking and GSS, IVA does not use direction information to separate the sound sources and works stand alone without prior knowledge of the sound source locations. In order to evaluate the separation performance of the algorithms, binary masking and GSS are equipped with the correct localization information to evaluate the pure separation performance of the algorithms.

After choosing a suitable separation algorithm, a third experiment is conducted to evaluate the separation performance in combination with the localization performance of the chosen algorithm.

Sound Source Separation: Experimental Results

The BSS EVAL toolbox [174] is a frequently utilized toolbox to evaluate source separation algorithms. The BSS EVAL toolbox is also used by the signal separation evaluation campaign [6] and is therefore regarded as an adequate means of judging the separation performance of the afore mentioned sound source separation algorithms. The BSS EVAL toolbox decomposes an estimated source signal into three signals, namely, s_{target}, e_{interf} and e_{artif}. The s_{target} can be described as the signal part that can be obtained by a convolved version of the original sound source. The e_{interf} can be explained by convolved versions of interfering original sound sources and e_{artif} is the signal part which can not originate from the original sound sources. The BSS EVAL toolbox then computes three measures to compare separation results, namely, the signal to distortion ratio (SDR), the signal to interference ratio (SIR) and the signal to artifact ratio (SAR). The SDR is computed by

$$\text{SDR} = 10 \log_{10} \frac{||s_{\text{target}}||^2}{||e_{\text{interf}}||^2 + ||e_{\text{artif}}|^2} \tag{3.12}$$

and is a measure of the amount of arbitrary distortions, i.e. interference from other sources and artificial noise in the sound signals after applying the separation algorithms. The SIR, given by

$$\text{SIR} = 10 \log_{10} \frac{||s_{\text{target}}||^2}{||e_{\text{interf}}||^2} \tag{3.13}$$

serves as a measure of interfering sound sources. Analogously, the SAR value, calculated by

$$\text{SAR} = 10 \log_{10} \frac{||s_{\text{target}}||^2 + ||e_{\text{interf}}||^2}{||e_{\text{artif}}||^2} \tag{3.14}$$

gives a judgment of artificial noise that is present in the separated signals. In other words, the higher the SIR value, the less interference of other sources are left in the separated sound signal and the higher the SAR value, the less artificial noise is introduced by the separation algorithm. Consequently, the higher the SDR, the less interferences and artificial noises are left in the separated signals.

Binary Masking

The separation performance of the binary masking approach is given in Tables 3.1, 3.2 and 3.3. The separation performance deteriorates if echoes are present in the recorded mixtures. This can be observed by comparing the anechoic audiolab experiments with the echoic experiments. Furthermore, the separation results depend on the number of sound sources in the mixture. The results in separating two simultaneously active sound sources are better than processing three simultaneously active sources.

The array geometry does not influence the separation performance significantly, which can be seen by comparing the separation results based on the recordings of Array 1, Array

2 and Array 3. The dependency of the echoes and of the number of sound sources can be described by Figure 3.1: the more echoes and the more sound sources the more overlap between the spectra of the single sound sources can be observed. Consequently, a mask that seeks to separate one sound source also contains overlapping snippets of other sound sources which causes lower SIR, SAR and SDR values in such a scenario.

Independent Vector Analysis

IVA demands at least as many microphone recordings as simultaneously active sound sources. Extensive experiments of my research team were conducted to evaluate the IVA method with different subsets of microphones and the subspace method [9] in different environments [168]. The experiments showed that the best results can be expected by utilizing all of the microphone array's microphones. Therefore, the presented results are based on IVA-separation that include eight microphone signals.

Tables 3.1, 3.2 and 3.3 give an overview of the separation performance that is achieved by IVA. Similar to binary masking, the separation performance is best for two simultaneously active sound sources in an anechoic environment. Separation in terms of SDR values for IVA separated mixture based on the Array 2 recordings are slightly better than for Array 1 recordings. Separation values based on IVA and Array 3 recordings are comparable to Array 1 recordings. The reason for the small separation differences between Array 2 and the other two arrays might be the shadowing effect of the concha applications of Array 2 leading to higher level differences of the microphone recordings which positively influence the IVA separation algorithm.

Geometric Source Separation

The separation performance of GSS for the different microphone arrays is presented in Tables 3.1, 3.2 and 3.3. According to IVA and binary masking separation, the best separation values are achieved in the anechoic environment with two simultaneously active speech sources. Between the different array configurations, there are no significant separation performance variations.

Binary Masking vs. IVA vs. GSS

According to the SIR, SAR and SDR values, binary masking can be excluded for further considerations since the separation performance by binary masking in the teleconference scenario is not competitive to IVA and GSS. The objective BSS EVAL toolbox evaluation results are further confirmed by listening to the separated results.

Regarding the SIR, SAR and SDR values, GSS outperforms IVA in the different array configurations and in both the anechoic and the echoic environment. However, one has to keep in mind that GSS uses extra localization information which was correctly provided for the sole comparison of the different separation algorithms.

For fair comparison of IVA and GSS, the real-world localization process for the GSS sound separation has to be included. Therefore, another experiment is conducted which applies GSS in combination with the SRP-PHAT localization algorithm. As seen in Table 3.4 the separation performance of the GSS decreases if real sound source localization replaces the ideal localization assumption. However, GSS is still competitive to IVA and even outperforms the IVA for Array 1. IVA has slight separation evaluation advantages for Array 2. For Array 3 there is no winner between IVA and GSS.

3.3. Concluding Remarks: Sound Source Separation

Extensive separation experiments show that the GSS separation outperforms the binary masking algorithm and IVA sound source separation in terms of SIR, SAR and SDR values for the three microphone arrangements and in the anechoic and the echoic condition, if correct localization values are assumed.

If the localization precondition for the GSS separation is fulfilled by realistic sound source localization that is done by the SRP-PHAT algorithm, the difference between GSS separation and IVA separation decreases. Consequently, there is no overall clear winner among IVA and GSS in terms of SIR, SAR and SDR values.

Concerning the teleconferencing scenario, I prefer the combination of GSS separation and SRP-PHAT sound localization since the localization results can be additionally used for the problem of speaker assignment. Furthermore, I decide to use the microphone array configuration Array 1 for further experiments since the SRP-PHAT localization performance in combination with Array 1 outperforms the other array configurations.

The BSS EVAL toolbox determined separation performance of SRP-PHAT in combination with GSS for Array 1 is better than using IVA in combination with Array 1 and the separation performance of SRP-PHAT in combination with GSS is competitive to the best IVA separation performances that are achieved with Array 2.

Algorithm	Anechoic			Echoic		
	SDR	**SIR**	**SAR**	**SDR**	**SIR**	**SAR**
Elevation: 10°: two sources						
GSS	**12.1**	**15.7**	**15.0**	**4.2**	12.1	5.3
IVA	5.3	11.6	8.5	-0.7	7.8	1.4
Binary Masking	3.2	14.6	3.9	-1.9	**12.4**	**5.5**
Elevation: 10°: three sources						
GSS	**9.8**	**13.4**	**12.9**	**2.9**	**9.8**	**4.5**
IVA	4.4	9.2	8.0	-1.7	4.9	1.4
Binary Masking	0.3	10.4	1.3	-3.7	8.4	-2.4
Elevation: 20°: two sources						
GSS	**10.8**	13.6	**14.5**	**4.7**	11.9	**6.0**
IVA	5.2	10.6	8.1	0.0	8.3	2.0
Binary Masking	3.0	**14.5**	3.6	-1.0	**12.8**	-0.4
Elevation: 20°: three sources						
GSS	**8.7**	**11.4**	**13.0**	**3.3**	**10.0**	**5.0**
IVA	3.7	8.6	7.7	-1.0	5.7	1.8
Binary Masking	0.2	10.5	1.2	-3.1	8.5	-1.8

Table 3.1.: Array 1: Comparison of the different sound separation approaches with recordings from the planar array [58]. All values are given in *dB*.

Algorithm	Anechoic			Echoic		
	SDR	**SIR**	**SAR**	**SDR**	**SIR**	**SAR**
Elevation: 10°: two sources						
GSS	**12.0**	15.2	**15.1**	**5.0**	12.2	**6.3**
IVA	6.9	14.6	9.4	0.8	9.5	2.9
Binary Masking	3.8	**15.6**	4.4	-0.4	**13.6**	0.2
Elevation: 10°: three sources						
GSS	**9.1**	**11.6**	**13.2**	**3.4**	**9.2**	**5.3**
IVA	5.5	11.1	8.6	-0.1	7.1	2.4
Binary Masking	0.6	10.9	1.6	-2.7	8.9	-1.5
Elevation: 20°: two sources						
GSS	**11.5**	**14.7**	**14.8**	**5.7**	**12.9**	**6.9**
IVA	6.0	11.9	8.6	1.3	9.7	3.3
Binary Masking	2.9	14.3	3.6	-0.6	12.5	0.1
Elevation: 20°: three sources						
GSS	**9.2**	11.7	**13.1**	**3.9**	**10.0**	**5.7**
IVA	6.8	**12.8**	9.4	1.3	9.2	3.2
Binary Masking	-0.2	9.9	1.0	-2.9	8.2	-1.5

Table 3.2.: Array 2: Comparison of the different sound separation approaches with recordings from the concha-microphone array [58]. All values are given in *dB*.

Algorithm	Anechoic			Echoic		
	SDR	**SIR**	**SAR**	**SDR**	**SIR**	**SAR**
Elevation: 10° *: two sources*						
GSS	**11.1**	**15.1**	**13.8**	**4.0**	**11.9**	**5.1**
IVA	6.0	13.7	8.7	-0.6	8.1	1.6
Binary Masking	2.8	14.0	3.6	-2.1	11.8	-1.3
Elevation: 10° *: three sources*						
GSS	**9.0**	**12.3**	**12.2**	**2.9**	**9.7**	**4.5**
IVA	4.0	9.1	8.0	-1.9	4.8	1.3
Binary Masking	-0.1	9.9	1.1	-4.1	7.6	-2.6
Elevation: 20° *: two sources*						
GSS	**10.9**	**14.6**	**13.6**	**4.7**	**12.4**	**5.9**
IVA	4.5	10.3	7.7	-0.2	8.2	2.1
Binary Masking	2.4	13.6	3.2	-1.6	11.7	-0.8
Elevation: 20° *: three sources*						
GSS	**8.9**	**12.1**	**12.2**	**3.3**	**10.1**	**4.9**
IVA	4.8	11.1	7.9	-0.1	8.0	2.0
Binary Masking	-0.6	9.6	0.7	-3.8	7.7	-2.3

Table 3.3.: Array 3: Comparison of the different sound separation approaches with recordings from the shield-microphone array [58]. All values are given in *dB*.

SRP and GSS	Anechoic			Echoic		
	SDR	**SIR**	**SAR**	**SDR**	**SIR**	**SAR**
Elevation: 10° *: two sources*						
Array 1	7.7	15.0	9.1	2.4	11.7	3.4
Array 2	5.8	14.3	6.8	1.9	11.5	2.8
Array 3	7.8	14.5	9.4	2.3	11.4	3.4
Elevation: 10° *: three sources*						
Array 1	2.2	11.3	3.3	-1.0	8.4	0.2
Array 2	0.5	8.8	2.0	-2.1	6.6	-0.3
Array 3	2.3	10.2	3.7	-1.3	7.4	0.3
Elevation: 20° *: two sources*						
Array 1	7.3	12.9	9.3	2.7	11.4	3.8
Array 2	5.8	13.7	6.9	2.6	12.3	3.4
Array 3	8.0	14.1	9.7	3.1	12.0	4.1
Elevation 20° *: three sources*						
Array 1	1.0	8.7	2.8	-1.7	7.4	-0.1
Array 2	0.5	9.0	2.0	-1.8	7.0	-0.1
Array 3	2.5	10.7	3.7	-0.7	8.4	0.7

Table 3.4.: Separation performance of the GSS algorithm fed with the localization data from the SRP-PHAT sound localization algorithm [58]. All values are given in *dB*.

4. Online Speaker Recognition

The identification of a conferee by the individual voice features can be considered as a useful advancement to assign conference participants to their individual transmission channels, compared to the assignment using the localization data. For example, the change of the position while conducting a conference can be properly handled by a speaker recognition system while the position change can not be detected by a localization based channel assignment.

The key requirements of a speaker recognition algorithm for the teleconferencing system are that the system is capable of online processing and text independent speaker identification. The online requirement means that the identification of the respective active speaker should be done as fast as possible to assign the speech contribution to the speaker's individual audio channel without audible delay. Since a system for conference applications has no prior knowledge about the user's speech contributions, the speaker recognition needs also to be text independent.

One way to identify an active speaker is to construct models by short term spectral features of each conference participant. A likelihood score between the model and an active speaker decides which particular person is speaking. The speaker dependent features should be robust against the conference participant's variability in mood or health condition during a meeting and among different conference meetings since it is extremely difficult to reach exactly the same emotional state or health state in different meetings [74]. In accordance with [85], the chosen speaker features should have a large variability between different speakers and a small variability given for the same conferee, even if the conferee is in different moods or a different status of health. Furthermore, the speaker dependent features should be robust against noise and should frequently appear in conference contributions.

The most common short term spectral features are the Mel frequency cepstrum coefficients (MFCCs) [125], which have been well proven and which are hard to beat in real world applications [85]. Besides the MFCCs there are different speaker dependent short term spectral features like Mel frequency discrete wavelet coefficients that apply a discrete wavelet transform instead of the MFCC's cosine transformation in order to improve recognition especially in noisy environments [14]. Spectral subband centroids [165] is another spectral feature that seeks to improve speaker recognition in noisy environments. To save computational costs, linear predictive cepstral coefficients [178] can be applied in place of MFCCs. Instead of computing the short term spectral features by the magnitude spectrum, modified group delay features are derived from the phase spectrum. In [65] it is claimed that group delay features allow for similar recognition performance as using MFCCs.

Besides short term spectral features, voice source features can be applied for speaker recognition. Voice source features are based on the assumption that the glottal source and the vocal tract source are independent of each other. Therefore, the vocal source features can be computed by filtering the recorded signal with an inverse vocal tract filter. The voice source features like glottal flow derivate waveform [118], wavelet octave coefficients of residues [186] or voice source cepstrum coefficients [60] can be used to improve short term spectral features.

Another group of features are the prosodic features like intonation patterns, speaking rate and speaking rhythm [14]. However, prosodic features are derived over a period of time which renders this approach useless for online speaker recognition in a teleconference situation. The same applies for high-level features, for example speaker dependent words like "hmm" or the usage of a speaker specific sequence of words in a sentence.

The speaker dependent features of a teleconference introduction round can be used to construct a speaker model for each conference participant. During the conferences the features, obtained by online feature extraction, are compared with the speaker model of each participant to recognize the active conferee. Approaches to construct the speaker models are vector quantization techniques [65] and support vector machines [29]. Another frequently used approach to obtain speaker models are Gaussian mixture models (GMMs) that can be considered as an extension of the vector quantization method [88].

4.1. Teleconference Online Speaker Recognition Algorithm

In my teleconferencing system, the speaker recognition task is based on GMMs due to the required low computational complexity of applying GMMs. A universal background model (UBM) in combination with maximum a posteriori (MAP) adaptation, proposed in [126], lead to a very fast generation of speaker-dependent GMMs, while only few training data is needed. In [56] it is shown that this system approach is capable of performing online speaker recognition.

MFCC Features are extracted from the incoming audio stream, constituting a likelihood score for every model and thus identifying the active conference participant. The recognizer is not only capable of online speaker recognition but also continuously improving recognition performance by online adaptation of the speaker models.

Preprocessing and MFCC Extraction

To extract a speech signal's spectral features, the incoming audio stream is divided into frames by a hamming window. A voice activity detection (VAD) locates frames that contain speech contributions by comparing the frame energy to a threshold. Silent frames are consequently discarded in a similar way to the segmentation method in [56].

Then, a sequence of feature vectors is generated that represents speaker-dependent information in every speech frame of the preprocessed signal. To represent the characteristics of individual voices, the MFCCs have proven meaningful as spectral features for speaker recognition tasks [125]. In our system approach we calculate MFCCs for each speech frame. In addition, we use the spectral frame energy as feature. The feature vector is extended by the respective first- and second-order delta regression coefficients to incorporate dynamic information.

Gaussian Mixture Models

In our system approach, Gaussian mixture models are considered to model different speakers. Let the feature vector $z \in \mathbb{R}^n$ be modelled as an n-dimensional random vector, then the Gaussian mixture density is defined by

$$p(z \mid \lambda) = \sum_{k=1}^{K} m_k \, \mathcal{N}(z \mid \mu_k, \Sigma_k), \tag{4.1}$$

where $\mathcal{N}(z \mid \mu_k, \Sigma_k)$ is a unimodal Gaussian density, parametrized by a mean vector μ_k, and a covariance matrix Σ_k. Therefore, the mixture density is a weighted linear combination of K Gaussian densities with mixture weights m_k, that satisfy the constraint

$$\sum_{k=1}^{K} m_k = 1 \ ; \quad 0 \leq m_k \leq 1. \tag{4.2}$$

Collectively, the parameters of the density model are denoted by $\lambda = \{m_k, \mu_k, \Sigma_k\}$. The GMM for speaker modeling was introduced in [127] and has proven to be efficient and effective for text-independent speaker recognition tasks.

Given a collection of training vectors, the model parameters λ are estimated using the iterative expectation-maximization (EM) algorithm [38]. The EM algorithm iteratively refines the model parameters to increase the likelihood of the model for the training data. The log-likelihood

$$ln \, p(Z \mid m, \mu, \Sigma) = \sum_{n=1}^{N} ln \, \{\sum_{k=1}^{K} m_k \, \mathcal{N}(z_n \mid \mu_k, \Sigma_k)\}. \tag{4.3}$$

provides a score, measuring the match between a collection of feature vectors Z of analyzed speech and speaker GMMs. A speaker is assigned to the analyzed data by picking the GMM with the highest score.

Universal Background Model and MAP Adaptation

The teleconferencing system approach uses a universal background model (UBM). A UBM is a single GMM that is trained on speech samples from a large number of representative speakers. The main advantage is that the UBM has to be trained only once, which can be computed in advance for a wide variety of possible speakers. Then, the specific speaker models for a conference are derived from this UBM by individually adapting it. This leads to a time efficient creation of speaker-dependent GMMs, benefiting the user comfort and practicability, and it has been shown that GMMs adapted from a well-trained UBM yield high speaker recognition rates [126].

The adaptation in our system is done by maximum a posteriori (MAP) estimation which is also known as Bayesian adaptation [42]. In order to create a new model for a certain speaker, the UBM is taken and adapted with the training data of this speaker. MAP adaptation is used to adapt only the means of the speaker GMM, which is saving computational cost and increases the speaker recognition performance [126].

Besides generating speaker models, we also use in the teleconferencing system MAP adaption for online improvement of already generated models, while the speaker recognition is running. Since the training material for building up models is limited and may not adequately characterize the range of conference conditions, speaker models can be improved by adapting them with already processed test data as shown in [56]. This online adaptation leads to more comprehensive models, mitigating the effects of changes in conference situation or speaker condition. The adaption of the speaker models in the teleconference system is only conducted if the localization data fits the recognition data to avoid an adaption with a falsely recognized conferee [161].

4.2. Online Speaker Recognition: Experiment

In this section, the speaker recognition is applied to the so-called AMI meeting corpus [4] to evaluate the performance of the online speaker recognition system with different sets of parameters. Furthermore, an offline speaker diarization system is applied to the meeting corpus to explore improvements that could be reached by processing the whole audio recording after the conference was held.

Online Speaker Recognition: Experimental Settings

Two scenarios are considered in the experiments. First, the online scenario, where meeting recordings are evaluated to test the parameter setting for the speaker recognition task. Second, the speaker diarization scenario where a MFCC and GMM based speaker diarization system [181], the winner of the 2007 NIST evaluation [55], is applied to the AMI-meeting corpus to get an impression for the upper possible bound of a speaker recognition system if online-constraint is not an issue.

The speaker recognition experiments are conducted with the AMI meeting corpus which is divided into different sub corpora. In the speaker recognition experiments, the Edinburgh meeting compilation (ES2009 to ES2016) is chosen as task for the speaker recognition system. Each meeting consists of four parts. One part is used to train the UBM and the remaining parts are used to evaluate the speaker recognition. Similar to an introduction round, a 10 s part of each conference participant is used to train the speaker models. The AMI-meeting corpus provides a ground truth for the diarization error rate (DER) calculation.

The DER is obtained by calculating the sum of different error components by

$$DER = \delta_{miss-error} + \delta_{false-alarm} + \delta_{speaker-error}, \tag{4.4}$$

where $\delta_{miss-error}$ describes the conference contributions that are not identified as speech by the system and $\delta_{false-alarm}$ is the detection of speech by the system within conference time slots where actually no speech contributions are existent. The $\delta_{speaker-error}$ denotes wrongly identified speakers. The lower the *DER*, the better the speaker recognition performance.

The most important parameters for the teleconferencing system's online speaker recognition are the

- frame length of the processed sound streams (t_f),
- number of utilized GMM components (n_{GMM}),
- number of used MFCCs (n_{MFCC}).

For more information about parametrization of the online speaker recognition approach, refer to an exhaustive evaluation of my research team in [91, 161].

Online Speaker Recognition: Experimental Results

In this section, the speaker recognition results for the AMI meeting corpus evaluation are presented. The first experiment investigates the influence of the frame length on the speaker recognition performance. The second and third experiments seek to quantify the benefits of using a high number of MFCCs and GMMs. Finally, a fourth experiment shows the difference between the online speaker recognition and an offline speaker diarization approach.

Influence of the frame length

In Table 4.1 it can be seen that the performance of the online speaker recognition strongly depends on the allowed frame length upon which the algorithm has to decide which conferee is talking. The frame length in this work can be considered as an upper boundary of the input signal for the speaker recognition system. It is worth mentioning that the actual frame length is slightly smaller, depending on the sample rate of the audio interface and the used block size. With the used block size of 1024 samples and a sample rate of 48 kHz, the actual frame length for an upper bound of 0.1 s is 0.085 s, which corresponds to a frame that consists of four blocks. For a frame length of 0.1 s a DER of 54.6 % is reached in meeting ES2009, whereas the DER for a frame length of 1.0 s leads to a DER

Meeting ES20xx	09	10	11	12	13	14	15	16
Frame length (t_f)								
0.1 s	54.6	68.6	67.4	52.1	63.8	54.4	61.2	61.8
0.25 s	40.4	59.0	57.5	40.3	52.0	44.5	50.4	50.5
0.5 s	33.2	53.4	50.7	33.4	44.6	38.9	45.0	43.3
1.0 s	**31.8**	**51.6**	**47.0**	**28.5**	**38.9**	**36.9**	**42.9**	**38.5**

Table 4.1.: Speaker recognition performance of the online speaker recognition system in terms of *DER* [%] (n_{GMM} = 128, n_{MFCC} = 12).

of 31.8 %. Obviously, the frame with a length of 1.0 s contains more speaker-dependent information than a frame with a length of 0.1 s, which improves the DER.

As seen in Table 4.1, there are also DER-variations between the meetings (ES2009..ES2016), which may be caused by quality differences of the headset conference recordings, varying conference participants with diverse voices and different meeting dynamics resulting in various overlaps of conference participants, which influence the speaker recognition system performance and consequently the DER.

Influence of the Number of GMM Components

Meeting ES20xx	09	10	11	12	13	14	15	16
Number of GMM components (n_{GMM})								
16	57.6	68.2	68.6	59.9	70.8	61.8	68.4	68.7
32	54.7	**65.3**	66.4	57.6	68.2	60.6	66.4	66.0
64	56.6	69.0	68.7	53.8	65.7	56.4	63.5	62.9
128	54.6	68.6	67.4	52.1	**63.8**	54.4	61.2	61.8
256	**48.1**	70.8	**62.5**	**47.4**	65.9	**50.9**	**58.8**	**57.1**

Table 4.2.: Speaker recognition performance of the online speaker recognition system in terms of *DER* [%] (t_f = 0.1 s, n_{MFCC} = 12).

The variation of DER for different numbers of GMM components are not as big as the DER variations concerning different frame lengths. As illustrated in Table 4.2, the online speaker recognition DER in meeting ES2009 for n_{GMM} = 16 is 57.6 %, whereas the DER for n_{GMM} = 256 is 48.1 %.

In meeting ES2010 the DER does not improve when using a higher number of GMM components, e.g., the DER for n_{GMM} = 16 is slightly better than the DER for n_{GMM} = 256. Except meeting ES2010 and ES2013 the DER is best when using n_{GMM} = 256. Thus, a higher number of GMM-components seems to allow for a more differentiated speaker model that improves online speaker recognition.

Influence of the Number of MFCCs

Meeting ES20xx	09	10	11	12	13	14	15	16
Number of MFCCs (n_{MFCC})								
8	52.3	72.6	65.9	54.6	67.9	53.3	62.4	61.7
12	54.6	68.6	67.3	52.1	**63.8**	54.4	61.2	61.8
16	**46.7**	68.9	63.4	47.2	66.5	**50.9**	59.4	57.7
20	**46.7**	**67.8**	**60.1**	**46.7**	66.2	51.0	**58.5**	**56.1**

Table 4.3.: Speaker recognition performance of the online speaker recognition system in terms of *DER* [%] (n_{GMM} = 128, t_f = 0.1 s).

Table 4.3 unveils slight DER variations for the online speaker recognition algorithm with different numbers of MFCCs. The DER differences between n_{MFCC} = 8 and n_{MFCC} = 20 for the meetings are between 2.3 % in ES2014 and 7.9 % in ES2012. According to the number of GMM components, the online speaker recognition performs better with a higher number of MFCCs that allows for a better representation of the conference participant's voice.

Offline Speaker Diarization

Meeting ES20xx	09	10	11	12	13	14	15	16
offline diarization	20.4	33.8	39.7	33.2	37.5	34.7	32.0	41.0

Table 4.4.: Speaker recognition performance of the online speaker diarization system [181] in terms of DER [%].

In this experiment, we apply a speaker diarization system [181] to the meeting corpus recordings. The speaker diarization algorithm segments and resegments the meeting recordings such that an optimal clustering of speech contributions and silence is possible. Furthermore, there is no frame length constraint, therefore, an online processing is not possible with the speaker diarization system.

The results of the speaker diarization system are illustrated in Table 4.4. It can clearly be seen that most of the achieved DERs are lower than the DERs of the proposed online speaker recognition system. However, it also can be observed that the differences of the DERs between the diarization system and the online recognition system are small if the online speaker recognition is allowed to work on frame lengths of t_f = 1.0 s. For meetings ES2012 and ES 2016 the online speaker recognition algorithm even outperforms the offline speaker diarization system and the DER differences are marginal for meeting ES2013 and ES2014.

4.3. Concluding Remarks: Online Speaker Recognition

Both, our online speaker recognition approach and the speaker diarization system are tested for the AMI meeting corpus recordings. The online approach's overall DERs are slightly worse than the DERs achieved with the diarization system, but the online speaker recognition approach is more appropriate to assign speech contributions to individual transmission channels in a teleconference scenario due to the fact that the diarization system requires the whole conference recording in advance.

It is worth mentioning that the high DERs of the online speaker recognition approach illustrated in Tables 4.2 and 4.3 can be explained by the small frame length that is used for the online speaker recognition experiment. However, the stand alone speaker recognition system does not fulfill the requirements for reliable channel assignment in a teleconference situation due to high DERs for frame lengths of $t_f = 0.1$ s. The small frame length, however, is essential due to the required low mouth to ear delay, which renders frame lengths of more than 0.1 s useless.

If we insist on a frame length of $t_f = 0.1$ s, the remaining parameters found by the AMI-meeting corpus experiments [161] for the online speaker recognition approach are as follows:

- Number of MFCCs $n_{MFCC} = 20$
- Number of GMM components $n_{GMM} = 256$

The itemized parameters are used in Chapter 5 for the channel assignment task.

5. Channel Assignment

In this chapter the findings of the sound source localization, separation and online speaker recognition, presented in Chapters 2, 3 and 4 are applied to assign the conferees in the conference room to individual transmission channels. In order to fulfill the low mouth to ear delay requirements, the different algorithms are restricted to a delay of maximum $t_d = 0.1\,\mathrm{s}$.

5.1. Channel Assignment Algorithms

In this section, three channel assignment algorithms are presented. The first algorithm is based on the online speaker recognition system, whereas the second algorithm uses the localization data to assign the active conference participant. The third algorithm seeks to combine the benefits of the sound source localization and the online speaker recognition.

Assignment by Speaker Recognition (ASR)

The ASR algorithm decides which conference participant is active by using the afore mentioned online speaker recognition system ($t_f = 0.085\,\mathrm{s}$, $n_{MFCC} = 20$, $n_{GMM} = 256$). The online speaker recognition system is fed with the GSS-separated sound sources as we suggested in [138]. The algorithmic signal processing delay is mainly caused by the block size of the inputs for the localization, separation and online speaker recognition unit, which is 1024 samples in this work. With a sample rate of $r_s = 48000\frac{samples}{s}$ the algorithmic signal processing delay for the localization and separation unit is $t_{d_{loc,sep}} = 0.021\,\mathrm{s}$, and for the speaker recognition system $t_{d_{rec}} = 0.085\,\mathrm{s}$. Consequently, the ASR algorithm has an algorithmic signal processing delay of $t_d = 0.106\,\mathrm{s}$.

The advantage of this algorithm is the detection of the active conference participants solely on the characteristics of the voice. Thus, the position of the conference attendants can vary during the conference.

Assignment by Localization (AL)

The second algorithm performs the assignment of the active speaker solely based on the localization of the respective positions of the conference participants around the conference table. The sound source localization and separation is done by the SRP-PHAT and the GSS algorithm, respectively, which were selected by numerous experiments as mentioned in Sections 2 and 3.

Due to the precise and robust performance results of the SRP-PHAT localization algorithm, this approach of assigning the conference participants to individual channels promises low DERs with a low algorithmic delay of $t_{d_{loc,sep}} = 0.021\,\text{s}$. One drawback of the AL assignment algorithm is the lack of the ability to compensate if conferees change their position during the conference since the AL algorithm does not take individual voice features into consideration.

Assignment by Localization, Controlled by Online Speaker Recognition (ALSR)

The third algorithm seeks to combine the advantage of fast and robust assignment by localization data and the usage of voice features to also add flexibility to position changes of the conference participants during the conference by parallel usage of the localization and online speaker recognition approach.

The assignment is done by comparing the detected localization of an active conference participant with the localization information obtained by an introduction round at the beginning of the conference. If the localized speaker position is within the corridor of an attendant, the sound segment is assigned to this individual transmission channel. At the same time, the voice features found in the sound frame are evaluated by the online speaker recognition algorithm. If there is a discrepancy between the localization information and the results of the speaker recognition, the conference participant has changed the position and the algorithm's knowledge about the position change is updated. My research team I provide more detailed information about the ALSR algorithm and the position change detection in [161] and [123]. The algorithmic mouth to ear delay of the ALSR algorithm is in accordance with the AL algorithm $t_{d_{loc,sep}} = 0.021\,\text{s}$, since the initial assignment is done by the AL algorithm.

5.2. Channel Assignment: Experiment

In order to investigate the performance of the three algorithms for channel assignment in the teleconference system, extensive real world experiments are conducted in an anechoic and an echoic environment.

Channel Assignment: Experimental Settings

Similar to the sound source localization experiment in Chapter 2 and the sound separation evaluation in Chapter 3, the channel assignment experiments are conducted in echoic and anechoic environments. Again, for the sake of reproducibility of the experiments in echoic and anechoic environments, we record speech contributions of eight male and four female conference participants. Each conferee has a speech contribution of five minutes which is spread over the whole conference. Additionally, we record for each conferee 10 s of extra speech that can be used to train the conferee's speaker models. The recorded conference

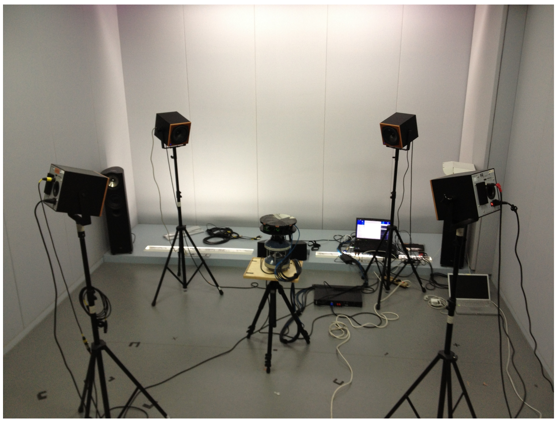

Figure 5.1.: Recording setting of the channel assignment experiment in the echoic environment

contributions are then arranged to a reproducible conference meeting that can be played back with loudspeakers placed around the conference table. The distance of the speakers from the center of the microphone array is 1.3 m and the elevation angle between the loudspeakers and the array is 20° which correspond to dimensions of a real conference meeting. Figure 5.1 illustrates the recording setup for the channel assignment experiment in echoic condition. The experiment environment and the experiment equipment is summarized in Table 2.1.

The conferences are arranged to fulfill the findings in [33], where it is found that during 26 different meetings, in 88% of the meeting time, only one conference participant is active, in 11% two speech contributions and in 1% three speech contributions are overlapping. We record the different conferences in the anechoic room and the echoic conference room. Furthermore, two combinations of possible conference participant's placement around the conference table are considered, refered to as placement 1 and placement 2 as illustrated in Figure 5.2.

In sum, 12 different conferences each with four participants are recorded. Each conference lasts about 19 min. The used training data to train the speaker models for the online speaker recognition system have a length of 10 s for each conference participant.

Channel Assigment: Experimental Results

The channel assignment by ASR, AL and ALSR is evaluated for the different conference recordings and placements of the conference participants. For each experiment, the DER is computed.

The achieved results of the respective algorithms can be observed in Table 5.1. The ASR algorithm that is doing the assignment solely on the online speaker recognition system results the worst DERs for the different conference situations. As already observed in Chapter 4, the small frame length of $t_f = 0.085$ s does not allow a reliable decision of the

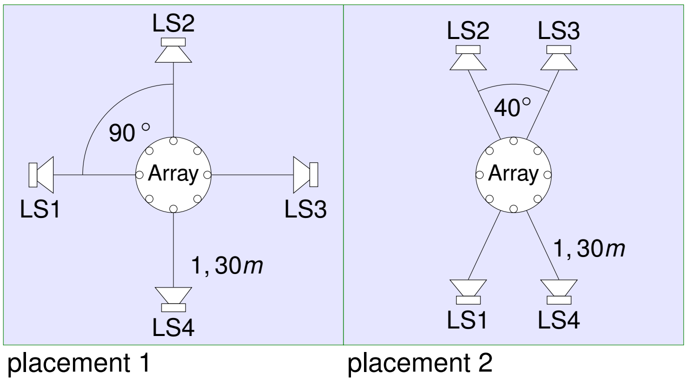

Figure 5.2.: Schematc overview of the conferee's placements in the experiments

Algorithm	Anechoic **placement 1**	 **placement 2**	Echoic **placement 1**	 **placement2**
Conference 1				
ASR	38.9	35.8	40.8	40.6
AL	17.1	18.1	20.0	21.1
ALSR	**17.0**	**16.9**	**19.7**	**20.7**
Conference 2				
ASR	38.1	34.2	37.5	35.6
AL	**20.7**	**17.4**	23.2	23.4
ALSR	20.8	**17.4**	**22.4**	**22.8**
Conference 3				
ASR	40.3	37.2	37.4	37.9
AL	**18.9**	**16.8**	23.4	24.4
ALSR	**18.9**	18.1	**22.7**	**23.2**

Table 5.1.: Speaker assignment results in terms of DER [%].

online speaker recognition system between the different conferees. However, the small frame length is a precondition to fulfill the delay requirement of the whole system. It is worth mentioning that the silent parts of the meeting usually also contribute to the DER due to the fact that the time slots without any active conference participant are assigned to an arbitrary audio channel. However, regarding the immersive playback, the assignment of silent parts has no influence to the listener. Therefore, the assignment of silent conference parts do not contribute to misleading playbacks of the falsely assigned channel or have any audible artifacts. If this fact is considered, the "true" DERs will be approximately 10 % lower than the DER values presented in Table 5.1.

Figure 5.3 shows an extract of a conference. The thin lines with the bright waveforms indicate the ground truth of the respective conference contributions. The dark lines in the ground truth streams of the conferees denote the decision of the ALSR algorithm which conference participant is active at the respective time instances. The dark waveforms, denoted by "Thomas assigned" , "Kathrin assigned", "Jonas assigned" and "Alex assigned" represent the audio signal outputs of the individual channels of the conferees that can be used to feed the immersive playback system. It can be seen, e.g. in the "Jonas assigned" that there are some falsely assigned time instances besides the correctly detected conference participant which can cause small peaks in the assigned channels. However, the assignment failures are not audible due to the very short duration of the misassignments.

The vertical gray line in Figure 5.3 indicates an exchange of positions between Jonas and Kathrin. It can be observed that initially, Jonas is falsely assigned to Kathrin's channel, since the new location of Jonas is Kathrin's former position. But the contradiction between the localization information and the voice features of Jonas are corrected by the speaker recognition algorithm such that Jonas is assigned correctly to his former channel after approximately four seconds.

5.3. Concluding Remarks: Channel Assignment

Three algorithms are tested for several realistic conference situations. The stand alone online speaker recognition approach (ASR) is not applicable in teleconferences due to worse assignment results which are mainly caused by the restricted algorithmic latency of the assignment task. The assignment by the localization information (AL) has a low algorithmic latency and low DERs. However, a position change of the conference participants is not detected by this approach. To overcome these problems, I suggest to combine the sound source localization assignment and the speaker recognition assignment (ALSR), that proved to achieve precise and robust assignment results within a realistic conferencing scenario. Furthermore, experiments show that the position change of conferees can be detected with the ALSR algorithm. So far, the ALSR algorithm is implemented and tested offline but a low algorithmic delay is an important prerequesite to adapt the ALSR algorithm to an online implementation. The algorithmic delay of the ALSR algorithm is $t_{d_{ALSR}} = 0.021\,\text{s}$ and in [138, 142], I found that a 3D-sound synthesis that is processed by a central conference server needs $t_{d_{trans}} = 0.178\,\text{s}$ for the transmission and the 3D HRTF-based sound synthesis. In sum, a complete delay of $t_{total} = 0.199\,\text{s}$ can be achieved for the complete conferencing system which is described as *very user satisfying* according to the ITU G.114 recommendation [77]. Due to the experimental results, I decide that the ALSR algorithm is applied in the developed conferencing system to assign the conference participants to their individual channel.

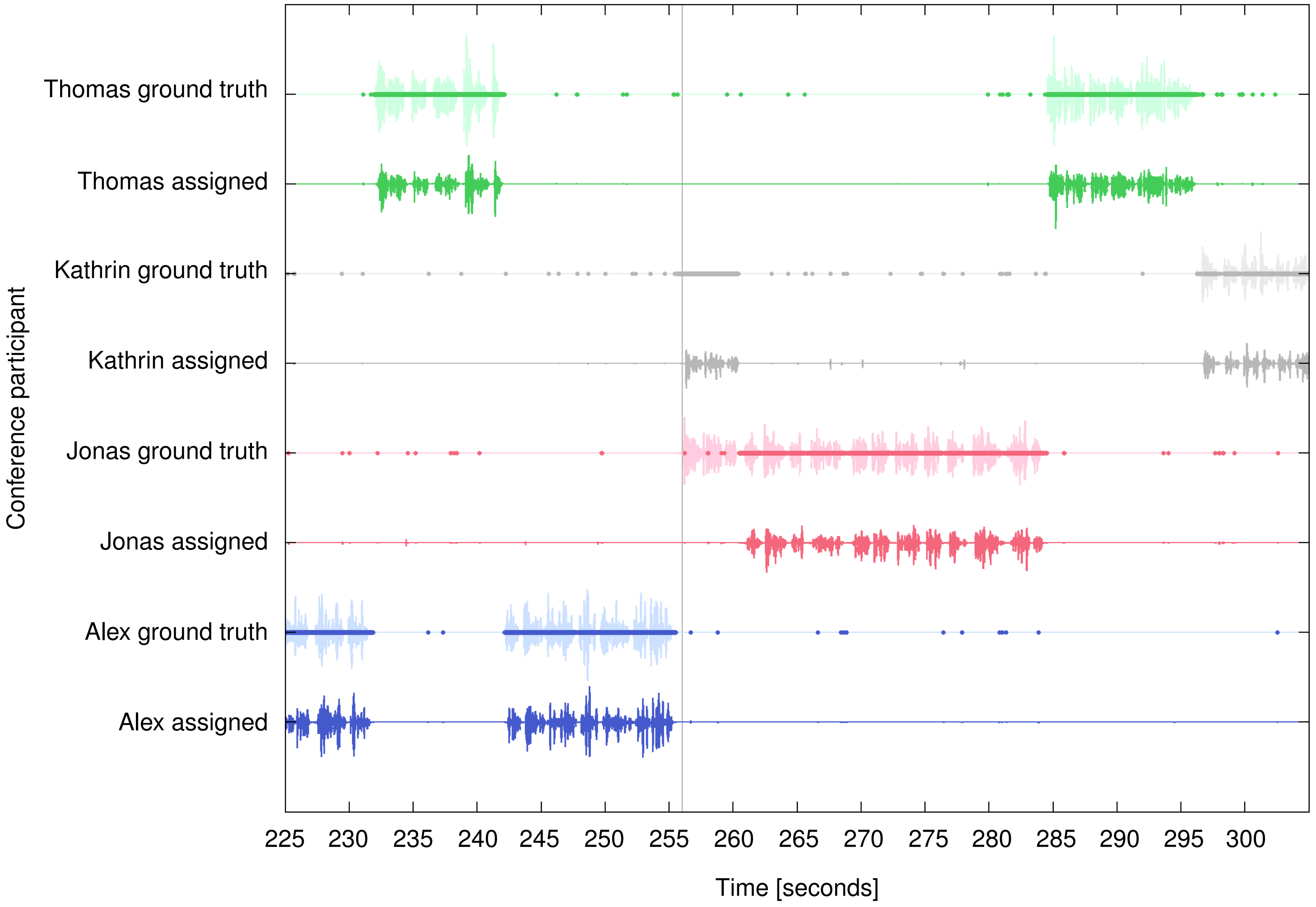

Figure 5.3.: Conference extract with four participants and a position change of two conferees, denoted by a gray vertical line [123]

Part II.

Immersive Playback

The second part of the thesis is about the immersive playback of the conference room contributions to a remote listener. My aim is to give the remote participant the acoustic impression of actually sitting at one conference table with all conference participants without being forced to use expensive extra hardware. In this work, I focus on 3D sound synthesis using head-related transfer functions (HRTFs).

HRTFs describe spectral changes of sound waves caused by diffraction and reflection off the human body, e.g. the head, shoulders, torso and ears [16]. In the last decades, HRTF-based techniques have become prominent in acoustic signal processing for various telepresence applications, e.g. binaural sound localization and synthesis [177] and binaural robotic sound source localization [44]. As each individual has, in general, a unique body geometry, the corresponding HRTFs are naturally different from person to person. Usually, HRTFs are obtained from recorded head-related impulse responses (HRIRs), which are the time domain representations of the HRTFs. 3D sound synthesis of the separated and assigned conference sound streams can be achieved by convolving a monophonic sound signal with HRIRs of the left and right ear that correspond to a certain direction. The convolved signal then can be played back via headphones or a specially adjusted loudspeaker assembly [179]. Regarding the teleconference scenario, headphone based playback seems to be the most appropriate approach of sound playback.

In order to achieve optimal sound synthesis for the headphone wearing remote conferee, I focus on differently obtained HRTF datasets:

- **KEMAR HRTF Dataset**:

  The KEMAR HRTF dataset describes a database that has been measured for a mannequin [54]. The database can be used for binaural sound synthesis without paying attention to the individual geometric features of the listener.

- **HRTF Selection Method**:

  The HRTF selection method describes an approach that selects the best possible HRTF data set among a HRTF database by playing back a sound that should move around a listener's head or should appear at a certain direction. The user then decides which data set produces the most immersive impression or the best direction accordance [81, 154]. This way, a data set of another person can be chosen for the sound synthesis by a procedure that is applicable in teleconference systems. This approach only allows for an approximation and does not generate an individual HRTF set with respect to the user's unique torso, head and pinna geometry.

- **HRTF Regression Method**:

  In contrast to the selection of a complete data set, regression methods can be utilized to generate an individual HRTF data set by using several acoustically measured HRTF training sets with the knowledge of the corresponding anthropometric data of the measured person. A regression model is constructed between the features of measured HRTFs and the anthropometry of the corresponding people. Finally, by knowing the anthropometry of a new person, one can generate its HRTFs by regression [135].

- **Individual Acoustic HRTF Measurement:**

  The most individual HRTFs for 3D sound synthesis applications, however, are obtained by acoustic measurement of the user. Usually, the acoustic measurement of HRTFs is done in an anechoic or semi-anechoic chamber.

The four alternatives of selecting a HRTF database describe a tradeoff between the degree of individualization and the effort (time, equipment) to be invested for the respective method. The most straightforward way to provide the remote listener HRTF-based sound synthesis is to use a standard HRTF dataset, e.g. the KEMAR dataset. With this approach the user does not have to adjust the playback system. The HRTF selection approach also does not require any extra hardware or acoustic measurement. The listener chooses among several datasets by a once-only procedure. This way, the degree of individualization is increased at the expense of time needed for the selection procedure. With the regression method, the degree of individualization is higher than for the selection method, because the listeners individual geometry is incorporated. Therefore, the anthropometric data has to be determined, i.e., an extra anthropometry measurement process is needed, making the regression method slightly more complex than the HRTF selection approach. Usually, the measurement of the required anthropometry requires much less effort than acoustically determining the HRIRs. Developments in the field of 3D scanners that I proposed in [137] or scanning software in combination with hardware, like the *Microsoft Kinect*, even improve the availability of anthropometric data for HRTF customization using regression.

The HRTF selection method, the HRTF regression method or the usage of KEMAR HRTFs are just approximations of the actual individual HRTFs of a listener. Therefore, the most individual HRTFs for 3D sound synthesis are achieved by acoustic measurement of the user, which is the most exhaustive way to generate a customized set of HRTFs. Usually, an anechoic or semi-anechoic chamber in combination with expensive hard- and software is needed for acoustic HRTF measurement. Furthermore, the listener has to be present for the whole measurement procedure, which can last for several hours, depending on the density of the spatial sampling grid and the chosen measurement approach.

In the second part of this thesis, I want to provide approaches to obtain HRTF datasets for 3D sound synthesis with respect to the geometric features of the remote listener of a teleconference. Different degrees of individualization are provided depending on the possibilities of the listener to access a recording environment or anthropometric data.

6. HRTF-Customization by Selection

The HRTF-customization by selection describes an approach to provide a listener with an individualized set of HRFTs for the sound synthesis without requiring any physical measurements. The listener himself chooses a set of HRTFs among different people's HRTFs within a listening test procedure. In this chapter, two related HRTF selection approaches are described and compared by preliminary listening tests.

Seeber-Fastl Method (SF)

In [154] a subjective selection method is presented where the authors suggest a two-stage selection method. In the first stage, a rough preselection reduces the number of possible eligible HRTF datasets from twelve to five. The second stage finally chooses the most fitting set of HRTFs by a more refined questionnaire.

In the first stage, the listeners are confronted with a test sound (five pulses of white noise) that is virtually synthesized to five positions in the frontal horizontal plane (-40°, -20°, 0°, +20°, +40°) by each of the twelve HRTF datasets. According to the best spatial perception, the listener preselects five datasets for the second stage.

In the second stage, the five top ranked HRTF sets of the first stage are used to again virtually synthesize the afore mentioned test sounds. Now, the listener should evaluate the datasets by refined criteria. The first criterion is that the perceived direction of the sounds should correspond to the synthesized sound sources. The second and third criteria ask the listeners if there are changes in elevation and distance during playback of the synthesized test signals. Ideally, such changes should not occur. The last criterion requires the synthesized sound source to be perceived at some distance and not in the head.

One key feature during this selection process is the self-determination of the user within one stage, meaning that the listener can choose the order of the possible HRTF datasets on his own. Also the user is allowed to listen to the datasets as often as required to make a descision.

DOMISO

Beside the SF approach, another similar approach, called determination method of optimum impulse-resonse by sound orientation (DOMISO) [81], seeks to select a proper set of HRTFs by a listening procedure.

The main difference to the afore mentioned approach is the selection by a Swiss-style tournament listening test where the listeners have no influence on the order of the played back test sounds. In each round, the listener has to choose between two test sounds and the tournament rules decide the final ranking of the corresponding HRTFs. In contrast to the SF approach, the DOMISO approach's test sounds (pink noise, duration: 1 s) are virtually synthesized around the listener in the horizontal plane.

6.1. Preliminary Listening Experiment

In this section, we apply the SF method and the DOMISO method to select a dataset from a number of presented HRTF datasets in order to decide which selection method is further utilized for the conferencing system.

Experimental Setting

In the experiment, the CIPIC database [3] is used for the HRTF selection methods. The database contains 35 human left and right ear HRIR tensors. We segmented the HRTF sets of the CIPIC database by the interaural time delay (ITD) into twelve groups. Each group consists of HRTFs with similar ITDs. One representative of each group is picked for the selection procedures. This way, we enable the listener to choose among HRTF datasets that cover the maximum of ITD variety within the dataset. The listening experiments are conducted in the institute's semi-anechoic room [145] to guarantee constant listening conditions.

We compare the ordering of the chosen datasets by the listeners for the different selection methods as well as the localization accuracy achieved with the respective favorite dataset. The localization tests are conducted by playing back test signals (white noise) virtually synthesized to 14 different directions around the listener in the elevation zero plane. Each direction is randomly played back five times. Consequently, there are in sum 210 test sounds played back for the SF method, the DOMISO method and for a KEMAR HRTF dataset that serves as non individualized reference.

Experimental Results

Two experiments with twelve test subjects are conducted. In the first experiment, the SF method is presented to the listener. In the second experiment, the HRTF selection is done according to the DOMISO approach. Finally, the Seeber-Fastl method and the DOMISO method are compared by direction localization tests with the listener's previously chosen favorite HRTF datasets. Besides the localization accuracy of the winner-database a ranking of the five favorite datasets for each listener is observed. The mean angular error (MAE) of the twelve test subjects for the KEMAR reference database is 37.43°, for the SF method selected winner 35.51° and for the DOMISO selection 33.49°. Therefore, averaged over the twelve test subjects, both selection methods slightly improve the proband's localization accuracy compared to the KEMAR dataset.

Exemplarily, I choose "Listener 10" to illustrate the sound localization results with the SF and DOMISO selected dataset. Out of the twelve eligible HRTF datasets, the proband has chosen the eleventh dataset for the DOMISO method and the tenth HRTF set for the SF method, meaning that the test subject prefers datasets with higher ITDs. The selection process lasts 961 s for the DOMISO method and 928 s for the SF method. According to the listening tests, both selection methods slightly improved the localization accuracy of the test subjects compared with the localization accuracy achieved with the KEMAR reference database.

Table 6.1 shows an overview of the results for the different selection approaches. Besides the mean localization error, the number of the respective winning dataset is given. Among the twelve probands, six listeners slightly improve the localization accuracy compared with the KEMAR-database for both selection methods. The SF method achieves for four test subjects better localization results than the DOMISO method. For eight people, the SF method selected dataset performs better than the reference KEMAR database. For the DOMISO selected database, ten probands achieve an improvement of the localization accuracy compared with the KEMAR database. Interestingly, three out of the twelve eligible HRTF databases were never chosen by any selection method and another two datasets were only chosen once. My research team and I provide further experiments and information for the HRTF selection approach in [57].

6.2. Concluding Remarks: HRTF Customization by Selection

According to preliminary listening tests, both selection methods can slightly improve the choice of the HRTF dataset in terms of the proband's sound localization accuracy. Based on the comparable results of the listening tests and due to the fact that less instruction is required, I decide the DOMISO method to be applied for further tests of the selection method's individualized HRTF datasets for the developed teleconferencing system. The method does not need any further equipment and can be done at an office workplace meaning that the remote listener in the conference scenario can autonomously conduct the individualization. However, the fact whether localization accuracy can be regarded as a sole element of judging the quality of an HRTF dataset in the context of teleconferencing is questionable. Therefore, a framework of comprehensive tests is conducted in Part III of this thesis.

	DOMISO	SF	KEMAR
Listener 1	11	6	-
MAE	**32.2**°	33.6°	38.3°
Listener 2	9	10	-
MAE	**39.0**°	42.2°	41.4°
Listener 3	4	11	-
MAE	45.3°	40.6°	**40.2**°
Listener 4	11	1	-
MAE	**14.4**°	36.7°	15.4°
Listener 5	4	5	-
MAE	**45.1**°	46.1°	47.4°
Listener 6	7	1	-
MAE	50.9°	**42.9**°	51.9°
Listener 7	11	11	-
MAE	20.6°	**18.0**°	25.6°
Listener 8	10	8	-
MAE	**28.9**°	36.8°	36.8°
Listener 9	1	9	-
MAE	**32.8**°	44.2°	40.1°
Listener 10	11	10	-
MAE	19.2°	**19.1**°	30.5°
Listener 11	5	10	-
MAE	43.3°	**39.8**°	41.2°
Listener 12	11	5	-
MAE	30.1°	**26.1**°	40.3°

Table 6.1.: Results of the preliminary listening tests to evaluate the DOMISO and the SF selection methods [57]. In the first line in the listener's results the number of the winning HRTF set is given and in the second line, the mean angular localization error (MAE) is presented.

7. HRTF Customization by Regression

Beside the HRTF selection (Chapter 6), another possibility to personalize HRTFs is to construct regression models between anthropometric data and direction-dependent features of HRTFs. The HRTF customization by using regression techniques allows to compute an individual set of HRTFs for each listener. Due to the generation of a new dataset for each user, I consider the regression method to deliver a higher degree of customization than the HRTF selection method. However, the regression method requires a set of anthropometric data of the respective listener, meaning that the effort needed for the regression method is slightly higher than the required effort for the selection method. The anthropometric data can be obtained by determining measurements of the listener with a caliper or, more advanced, with a 3D laser scanner as I proposed in [137].

There are already research efforts in customization of HRTFs, which aim to estimate HRTFs based only on geometric information of the listener without acoustically measuring their HRIRs [69, 117]. Such a process requires usually a collection of HRTF datasets of various subjects, which result in huge amount of data. Since the pioneering work [89], principal component analysis (PCA) has become a popular tool for HRTF reduction [111]. The application of PCA in HRTF customization [69, 117], which reduces the dimension of the original dataset before customization, has demonstrated promising performance. A collection of different proband's HRTFs can be considered as a three-way data array, whose three directions represent subject, location and frequency, respectively. Applying PCA to HRTF datasets requires in general a vectorization process of the original dataset. As a consequence, some useful information within the structure of the HRTF dataset might be disregarded. To avoid such limits, the so-called tensor singular value decomposition (T-SVD) method, which was originally introduced in the community of multiway array analysis [93], has been recently applied into HRTF customization successfully [59].

In the community of image processing, two recent techniques of multiway array analysis are proposed in competition with the standard PCA. The two dimensional PCA (2DPCA), originally a direct generalization of PCA for image analysis, and the so-called generalized low rank approximations of matrices (GLRAM) [183], a further generalized form of 2DPCA. I have demonstrated that GLRAM and T-SVD outperform the standard PCA in the task of dimensionality reduction of HRTFs [140], which can be considered as a sign that the methods successfully detect the direction dependent features in the HRTF datasets.

In this chapter, we study GLRAM, 2DPCA and Tensor-SVD methods for the purpose of customizing HRTF datasets and compare their performance with the standard PCA. Furthermore, partial least squares regression (PLSR) [180] is applied to construct a customized HRTF dataset and PLSR results are compared to the afore mentioned multiway principal components regression methods.

7.1. HRTF Customization

Given a set of measured HRTFs of different people, a multiple linear regression seeks to match a set of anthropometric parameters to the characteristics of the individual's transfer functions. In general, a collection of HRTFs can be represented as a three-way array $\mathcal{H} \in \mathbb{R}^{N_d \times N_f \times N_p}$, where the dimensions N_d is the spatial resolution of directions, N_f the frequency sample size and N_p is the number of people in the training dataset. For the sake of simplicity, we adapt a Matlab-style notation in this section, e.g., $\mathcal{H}(i,j,k) \in \mathbb{R}$ is denoted the (i,j,k)-th entry of $\mathcal{H}$, $\mathcal{H}(l,m,:) \in \mathbb{R}^{N_p}$ the vector with a fixed pair of (l,m) of $\mathcal{H}$ and $\mathcal{H}(l,:,:) \in \mathbb{R}^{N_f \times N_p}$ the l-th slide (matrix) of $\mathcal{H}$ along the direction-dimension.

In order to receive only the direction dependent information between the different individuals, the mean of the subject's average log-HRTFs is subtracted from each log-HRTF [89]. It results in an interindividual direction transfer functions between the subjects, denoted by $\mathcal{D} \in \mathbb{R}^{N_d \times N_f \times N_p}$, whose (i,j,k)-th entry is computed by

$$\mathcal{D}(i,j,k) = 20\log_{10}|\mathcal{H}(i,j,k)| - \frac{1}{N_p}\sum_{k=1}^{N_p} 20\log_{10}|\mathcal{H}(i,j,k)|\,. \tag{7.1}$$

An idea of customizing unknown HRTFs is to first extract certain direction dependent main features out of the directional transfer functions $\mathcal{D}$, then to construct a multiple linear regression model between anthropometric features of subjects and the extracted directional dependent features. Let $K = [k_1, \dots, k_{N_p}] \in \mathbb{R}^{r_p \times N_p}$ be a set of r_p chosen directional dependent features and $O = [o_1, \dots, o_{N_p}] \in \mathbb{R}^{N_o \times N_p}$ be a collection of N_o anthropometric features of test subjects. For the k-th subject, a multiple linear regression model between anthropometric parameters and direction dependent features can be constructed as

$$k_k = B\tilde{o}_k + \epsilon, \tag{7.2}$$

where $B \in \mathbb{R}^{r_p \times (N_o+1)}$, $\tilde{o}_k = [1 \ o_k^\top]^\top \in \mathbb{R}^{N_o+1}$, and $\epsilon \in \mathbb{R}^{r_p}$ is the estimation error vector. Let us denote $\mathbf{1} \in \mathbb{R}^{N_p}$ the vector with all entries equal to one, and construct $\tilde{O} = [\mathbf{1} \ O^\top] \in \mathbb{R}^{N_p \times (N_o+1)}$. It is known that N_p is usually greater than N_o. We assume that matrix $\tilde{O}$ is full rank. Then, in terms of minimization of the error ϵ by least squares method, the regression coefficient matrix B in model (7.2) is computed by

$$B = K\tilde{O}(\tilde{O}^\top\tilde{O})^{-1}. \tag{7.3}$$

Finally, given the anthropometric data $o_{new} \in \mathbb{R}^{N_o}$ of a person not in the training set, its transfer function features k_{new} can be constructed by

$$k_{new} = B\tilde{o}_{new} \in \mathbb{R}^{r_p}, \tag{7.4}$$

where $\tilde{o}_{new} = [1\ o_{new}^\top]^\top \in \mathbb{R}^{N_o+1}$. In this work, we choose the set of anthropometric parameters for multilinear regression in accordance with [69], where anthropometric parameters are selected by applying correlation analysis.

7.2. HRTF Customization Methods

This section briefly overviews three techniques of feature extraction methods for the dataset $\mathcal{D}$, namely, 2DPCA, GLRAM and Tensor-SVD. Moreover, the PLSR method is briefly described.

Customization Using 2DPCA

Similar to the popular approach of customizing HRTFs by using PCA, 2DPCA based HRTF customization can be described as follows. First of all, the so-called scatter matrix $S_p \in \mathbb{R}^{N_p \times N_p}$, given by

$$S_p = \frac{1}{N_d} \sum_{i=1}^{N_d} \mathcal{D}(i,:,:)^\top \mathcal{D}(i,:,:), \tag{7.5}$$

is calculated. Then r_p eigenvectors $K = [k_1, \dots, k_{r_p}] \in \mathbb{R}^{N_p \times r_p}$ corresponding to the r_p largest eigenvalues are computed. The so-called principal components of 2DPCA for the i-th slides of $\mathcal{D}$ are calculated by

$$\widehat{\mathcal{D}}(i,:,:) = \mathcal{D}(i,:,:)K. \tag{7.6}$$

The direction dependent regression coefficient matrix B is then constructed as given in Equation (7.3). A set of customized direction transfer functions $D_{new} \in \mathbb{R}^{N_d \times N_f}$ for an unknown person is obtained with its i-th slide given by

$$D_{new}(i,:) = \widehat{\mathcal{D}}(i,:,:)k_{new}^\top, \tag{7.7}$$

where k_{new} is computed in accordance with Equation (7.4). I refer to [82] for further information on PCA and to [182] for further discussions on 2DPCA.

Customization Using Tensor-SVD

Unlike customization using PCA, Tensor-SVD keeps the structure of the original 3D dataset intact and computes the customized dataset for every direction at once. Given a dataset $\mathcal{D} \in \mathbb{R}^{N_d \times N_f \times N_p}$, Tensor-SVD computes its best multilinear $rank - (r_d, r_f, r_p)$ approximation $\widehat{\mathcal{D}} \in \mathbb{R}^{N_d \times N_f \times N_p}$ [93]. The $rank - (r_d, r_f, r_p)$ tensor $\widehat{\mathcal{D}}$ can be decomposed as a *trilinear* multiplication of a $rank - (r_d, r_f, r_p)$ core tensor $\mathcal{C} \in \mathbb{R}^{r_d \times r_f \times r_p}$ with three full-rank matrices $U = (u_{ij}) \in \mathbb{R}^{N_d \times r_d}$, $V = (v_{ij}) \in \mathbb{R}^{N_f \times r_f}$ and $K = (k_{ij}) \in \mathbb{R}^{N_p \times r_p}$, which is defined by

$$\widehat{\mathcal{D}} = (U, V, K) \cdot \mathcal{C}, \tag{7.8}$$

where the (i, j, k)-th entry of $\widehat{\mathcal{D}}$ is computed by

$$\widehat{\mathcal{D}}(i, j, k) = \sum_{\alpha=1}^{r_d} \sum_{\beta=1}^{r_f} \sum_{\gamma=1}^{r_p} u_{i\alpha} v_{j\beta} k_{k\gamma} \mathcal{C}(\alpha, \beta, \gamma). \tag{7.9}$$

Finally, with the regression model built in (7.3) and (7.4), a new set of direction transfer functions can be retrieved by

$$D_{new} = (U, V, k_{new}^{\top}) \cdot \mathcal{C} \in \mathbb{R}^{N_d \times N_f} \tag{7.10}$$

Refer to [148] for Tensor-SVD algorithms and further discussions.

Customization Using GLRAM

Similar to Tensor-SVD, GLRAM methods do not require the destruction of 3D tensors. Instead of reducing the dataset $\mathcal{D}$ along all three directions as Tensor-SVD, GLRAM methods work with two pre-selected directions of a 3D data array. Given a dataset $\mathcal{D} \in \mathbb{R}^{N_d \times N_f \times N_p}$, the task of GLRAM is to approximate matrices $\mathcal{D}(:, i, :)$, for $i = 1, \dots, N_f$ of $\mathcal{D}$ along the second direction by a set of low rank matrices $\{UG_i K^{\top}\} \subset \mathbb{R}^{N_d \times N_p}$, for $i = 1, \dots, N_f$, where the matrices $U \in \mathbb{R}^{N_d \times r_d}$ and $K \in \mathbb{R}^{N_p \times r_p}$ are of full rank.

Similar to the 2DPCA and the Tensor-SVD method, a new set of direction transfer functions can be retrieved by

$$D_{new}(:, i, :) = U\mathcal{G}(:, i, :)k_{new}. \tag{7.11}$$

Further details on GLRAM algorithms can be found in [183].

Customization Using PLSR

While PCA, 2DPCA, GLRAM and TSVD based decomposition of the HRTFs seek to minimize the reconstruction error of the datasets, PLS regression method seeks to find regression weights with respect to the covariance of $\mathcal{D}$ and O [180]. Simultaneous to the multiple linear regression model of Equation (7.2) the idea of customizing unknown HRTFs by the use of PLSR is based on the assumption to express the direction transfer functions by $d_k = Q\tilde{o}_k + \epsilon$.

According to [180] the PLSR finds variables T that are predictors for $\mathcal{D}$, which also describe O meaning that $\mathcal{D}$ and O are partly modelled by the same features. The anthropometric data O and the transfer function for each direction $\mathcal{D}(i, :, :) \in \mathbb{R}^{N_f \times N_p}$ can be decomposed to $O = TP^{\top}$ and $\mathcal{D}(i, :, :) = IC^{\top}$, where T, I are the features of O and $\mathcal{D}(i, :, :)$, respectively. P and C are the corresponding weights. Assuming the features of O partly model $\mathcal{D}(i, :, :)$, the direction transfer function for each direction i can be described by $\mathcal{D}(i, :, :) = TC^{\top}$. With $T = OK^{\top}$, $\mathcal{D}(i, :, :)$ is defined by

$$\mathcal{D}(i, :, :) = OK^{\top}C^{\top} + F = OQ + F, \tag{7.12}$$

where F is the estimation error matrix.

Finally, the regression coefficient matrix can be computed by

$$Q = K^\top C^\top \tag{7.13}$$

and given the anthropometric data $o_{new} \in \mathbb{R}^{N_o}$ of a person not in the training set, its transfer function $D_{new} \in \mathbb{R}^{N_d \times N_f}$ can be constructed by

$$D_{new}(i, :) = Q\tilde{o}_{new}. \tag{7.14}$$

Refer to [108] and [180] for further information about PLSR.

7.3. HRTF Customization: Experiment

In this section, we apply PCA, 2DPCA, GLRAM, Tensor-SVD and PLSR to compute the individual HRTFs with regression. The performance of the different HRTF customization approaches is investigated and discussed.

Experimental Setting

In the experiment, the CIPIC database [3] is used for the HRTF customization application. The database contains 35 human HRIR tensors with the corresponding anthropometric data for both left and right ears. The CIPIC HRIRs are recorded in spatial resolution of $N_d = 1250$ points ($N_e = 50$ in elevation and $N_a = 25$ in azimuth), spaced uniformely around the head, with $N_t = 200$ time samples. To obtain the HRTFs, the discrete fourier transformation (DFT) was applied on each HRIR.

We use cross validation to compare the different methods. The cross validation method takes the direction transfer functions $\mathcal{D}$ of the people within the CIPIC dataset, together with their anthropometry as a training set to conduct the customization by regression, as explained in Section 7.1. The person to be reconstructed is not part of the training set. The customization procedure is repeated for each person of the training set in order to compare the regression performance for each subject.

For the regression model (7.2), we select the anthropometric parameters in accordance to [69]. It is demonstrated that eight selected parameters out of 27 from the original CIPIC database cover most of the variance of the dataset and provide good regression performance with minimal measurement efforts for the anthropometric data. These eight parameters are: head width, head depth, shoulder width, cavum concha height, cavum concha width, fossa height, pinna height and pinna width. The feature extraction parameters for the different regression methods are chosen based on exhaustive simulations in [92].

Experimental Results

In each experiment, we construct a new set of HRTFs for the person not in the training set with one of the introduced feature extraction methods. To investigate the performance of the different feature extraction approaches in a HRTF customization application, the spectral distortion for every angle over the whole frequency spectrum is computed. The spectral distortion SD is defined by

$$\mathrm{SD} = \sqrt{\frac{1}{N_f}\sum_{i=1}^{N_f}\left(20\log_{10}\frac{|H_i|}{|H_{new_i}|}\right)^2}, \tag{7.15}$$

where H_i is the magnitude of the CIPIC-measured HRTF in the dataset and H_{new_i} is the magnitude of the HRTF constructed via regression at the i-th frequency. N_f is the number of frequency samples for each HRTF (200 in this case).

First of all, PCA with $r_d = 10$ dominant eigenvectors is applied to the task. We use that as a reference for comparison with the other three multilinear methods. Table 7.1 summarizes the average spectral distortion values for the estimation of the left ear HRTFs. It can be seen that the customization procedure leads to different spectral distortion values from subject to subject. For subject 30 and subject 33, the estimation of the HRTF works quite well in comparison with subject 29. This might be caused by the possibility, that there exists a subject in the training set that is physically similar to these two subjects. Furthermore, it indicates that the precise determination of anthropometric parameters as well as the measurement of HRIRs is sensitive to many other parameters, e.g. the placement of the microphones or head movements during the measurement process. Such inconsistencies in the training set can consequently lead to different estimation values due to varying precisions of the regression models.

The direction dependency of the estimated HRTFs can be seen in Figure 7.1. For different directions, the estimation quality of the individual HRTFs slightly varies in terms of spectral distortion. Extractions of the direction dependent features with PCA disregard similarities between neighbouring angles. To also take the 3D structure of the dataset into account, 2DPCA is applied also using $r_d = 10$ eigenvectors. SD results of HRTF customizations at two particular planes, shown in Figure 7.1, indicate that 2DPCA extracted features lead to a better estimation of the HRTFs than PCA. Lower SD values suggest that the log-magnitude response of the estimated HRTF is closer to the CIPIC-measured one than the PCA estimated one.

Finally, GLRAM ($r_p = 10$, $r_f = 200$, $r_d = 100$) and Tensor-SVD ($r_p = 36$, $r_f = 36$, $r_d = 36$) are applied to estimate the individual HRTFs of the test subjects. The achieved SD results are similar to the 2DPCA regression results . Using PLSR does not further improve the regression SD results compared to 2DPCA, GLRAM and Tensor-SVD methods. My research team and I provide further information about HRTF regression in [92, 135, 143].

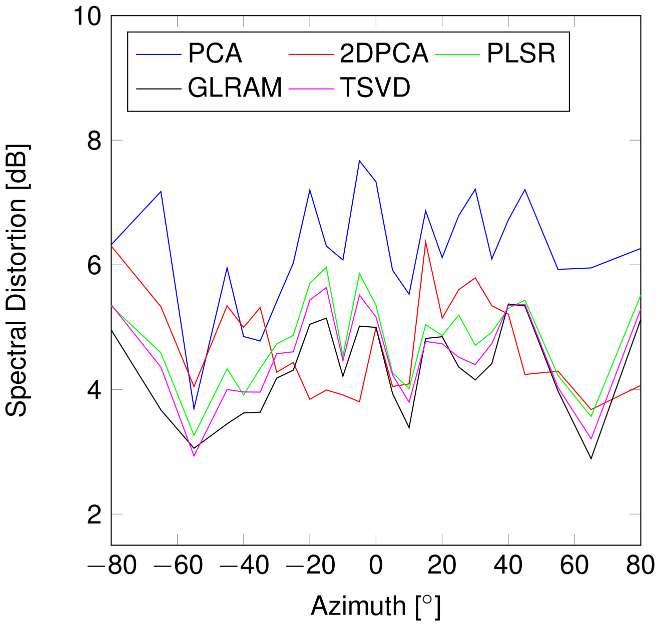

Figure 7.1.: SD values for Subject 30 in the horizontal plane

7.4. Concluding Remarks: HRTF Customiation by Regression

Our experiments demonstrate that multiway regression models, namely 2DPCA, Tensor-SVD and GLRAM outperform the standard PCA approach with respect to the spectral distortion values. SD values achieved by using PLSR are better than the PCA achieved SDs but inferior to the multiway array methods. The regression methods do not require individual acoustic measurements to obtain a customized dataset, however, the result of the regression method is strongly dependent on the measurement accuracy of the training set and the determination of the anthropometric data. Due to the fact that we can avoid acoustic measurement for HRTF customization, the regression methods seem to offer a promising approach in the teleconferencing scenario to improve immersive playback of conference contributions. Therefore, I choose a subset of the regression approaches (2DPCA, GLRAM, PLSR) to be evaluated in the listening tests in Part III of this thesis.

Person	PCA	2DPCA	GLRAM	TSVD	PLSR
Subject 1	6.86	5.68	5.34	5.72	6.27
Subject 2	6.81	5.50	5.37	5.96	6.36
Subject 3	7.95	6.77	6.54	6.70	7.01
Subject 4	6.88	5.66	5.55	5.86	6.26
Subject 5	6.96	5.48	5.37	5.62	5.86
Subject 6	7.29	6.17	5.95	6.40	6.85
Subject 7	7.81	6.54	6.22	6.57	6.96
Subject 8	6.60	5.61	5.39	5.70	5.96
Subject 9	7.00	5.93	5.85	5.85	5.92
Subject 10	6.89	5.60	5.52	5.80	6.14
Subject 11	8.74	6.72	6.49	7.08	7.73
Subject 12	6.77	5.65	5.56	5.83	6.04
Subject 13	6.81	5.36	5.31	5.36	5.47
Subject 14	7.00	5.94	5.89	6.23	6.53
Subject 15	7.27	6.13	6.09	6.18	6.29
Subject 16	7.79	6.74	6.66	6.70	6.85
Subject 17	6.65	5.75	5.60	5.69	5.95
Subject 18	6.61	5.42	5.31	5.44	5.68
Subject 19	6.17	4.92	4.75	4.92	5.21
Subject 20	7.12	5.75	5.60	5.81	6.12
Subject 21	7.48	7.14	7.12	7.28	7.49
Subject 22	7.26	5.73	5.65	5.76	6.07
Subject 23	7.08	5.90	5.81	6.12	6.45
Subject 24	6.33	4.98	4.96	5.12	5.34
Subject 25	6.52	5.74	5.69	5.85	6.11
Subject 26	9.10	7.44	7.41	7.71	7.98
Subject 27	6.54	5.73	5.61	5.71	5.85
Subject 28	7.41	6.01	5.99	6.04	6.14
Subject 29	7.54	7.18	7.18	7.20	7.24
Subject 30	5.71	4.39	4.36	4.48	4.59
Subject 31	7.06	5.68	5.41	5.87	6.41
Subject 32	6.22	5.25	5.15	5.19	5.41
Subject 33	6.17	5.04	4.89	5.20	5.45
Subject 34	7.47	6.13	6.01	6.38	6.66
Subject 35	6.49	5.33	5.18	5.58	5.91
mean	7.04	5.86	5.74	5.97	6.25

Table 7.1.: Average spectral distortion (SD) values for the different people in terms of dB.

8. HRTF Measurement

The selection methods presented in Chapter 6 and the regression methods described in Chapter 7 offer the chance to efficiently obtain individual HRTF datasets for the teleconferencing system. However, the most individual HRTF-based sound synthesis can be achieved by acoustically measuring the subjects actual head-related transfer function. In order to enable a remote listener to use an individual acoustically measured set of HRTFs, the user usually has to attend a cumbersome measurement procedure in advance.

There are various HRTF measuring techniques that compute the HRTF by recording the excitation signal (e.g. maximum length sequences, sine sweeps) for each spatial sampling point. One major problem is the huge amount of time that is needed to generate an individual, dense HRTF data set that is sufficient for immersive 3D sound synthesis, e.g. in order to synthesize moving sound sources or to enable the use of head tracking for dynamic sound synthesis. Another problem is that the person to be measured must not move during the recording procedure. This can be difficult if the subject is sitting on a turntable that accelerates and stops between the recordings. There are attempts to ensure good-quality HRTF databases by detecting head movements during the measurements via optical tracking systems [72]. However, the recordings during which the user moved, have to be repeated which in turn prolongs the recording procedure. To tackle these two drawbacks of conventional HRTF measurement approaches, a recent approach uses normalized least mean square (NLMS) adaptive filters to compute the HRIRs from the recorded excitation signals. Adaptive filters [66] are capable of estimating impulse responses and are able to handle defined movements of the subjects during the recording procedure. Therefore, it is possible to make measurements with continuous rotation, which reduces the recording time independently from the desired spatial resolution. With the NLMS method, it is possible to generate continuous HRTFs in azimuth without interpolation [46]. A generalization to multiple-elevation continuous-azimuth HRTF acquisition was also reported in [45].

The NLMS method could also be beneficial for loudspeaker-based 3D sound synthesis applications using crosstalk cancellation and head-tracking [159] to efficiently measure the required dense transfer function database.

In this chapter I study the well-established static HRTF measurement approaches using maximum length sequences (MLS) and sine sweeps and compare the methods with a HRTF estimation approach using normalized least mean square (NLMS) adaptive filters under the same lab conditions [141]. The different approaches are implemented and experimentally compared by objective and subjective evaluation.

8.1. HRIR Measurement Approaches

In this section, an overview of HRIR estimation approaches, namely, maximum length sequences (MLS), exponential sine sweeps, and NLMS adaptive filters is presented.

Maximum Length Sequence (MLS)

For the MLS-method [131, 152], a pseudo-random excitation signal with an impulse-like autocorrelation is used. The amplitude of the impulse depends on the variance and length of the excitation signal. The head-related impulse response h then can be computed by

$$h = \frac{g \star x}{\sum\limits_{m=-\infty}^{\infty} g[m]^2}, \tag{8.1}$$

where g is the excitation signal, x is the signal recorded with the ear microphones and $\star$ denotes the cross-correlation operation. One major advantage of the MLS method is that it is robust against transient noise because the energy of the disturbance is uniformly distributed along the impulse response [84]. We use a periodic MLS excitation signal which results in a periodic response of the system and average the result over two repetitions.

Exponential Sine Sweeps

The exponential sine sweeps method [48] avoids deconvolution instability by mathematically computing the inverse of the excitation signal. An exponential sine sweep excitation signal with a length of T which is covering the bandwidth from f_1 to f_2 can be computed by

$$g(t) = \sin\left[2\pi f_1 \frac{T}{ln(\frac{f_2}{f_1})} \exp\left(\frac{t}{T} ln\left(\frac{f_2}{f_1}\right) - 1\right)\right]. \tag{8.2}$$

The main advantage of the exponential sine sweeps method is the fact that one can avoid inversion instabilities by computing the inverse of the excitation signal by

$$g_{inv}(t) = g(T - t)\exp\left(\frac{-t}{T}\ln\left(\frac{f_2}{f_1}\right)\right) \tag{8.3}$$

and the impulse response h is obtained by $h = x * g_{inv}$, where $*$ denotes the convolution operation. Furthermore, exponential sine sweeps consist of only one frequency at one time instance, therefore non-linear distortions can be identified and removed.

HRIR Measurement Using NLMS-Type Adaptive Filters

In order to determine individual HRIRs with the MLS and exponential sine sweeps methods, it is necessary to predefine the resolution of the desired HRIR grid.

Sound source	KS digital C5 tiny
KEMAR microphones with IEC 60711 Coupler	GRAS-Type 40AG, RA0045
Preamp	GRAS-Type 26AC
Sound card	RME Multiface II
Lab dimensions	4.7 m x 3.7 m x 2.84 m
Lab noise level A-weighted	16.3 dB
Lab reverberation time t_{60}	0.08 s

Table 8.1.: Information about the equipment used in the HRIR measurement experiment.

Using the NLMS-type adaptive filters approach, it is possible to define the resolution of the HRIR grid after the measurement process. The basic idea of this approach is to track the system's true impulse response vector h by the estimated impulse response vector $\hat{h}$.

First, a prediction $\hat{x}[k]$ of the in-ear recordings $x[k]$ at discrete time k is computed by convolving the excitation signal $g[k]$ with the estimated impulse response $\hat{h}[\theta_k] \in \mathbb{R}^N$ as $\hat{x}[k] = g^T[k] \cdot \hat{h}[\theta_k]$, where N describes the length of the impulse response and $g[k] \in \mathbb{R}^N$ is defined as the section of g containing the current and the $N-1$ preceding samples of g at time instance k. The prediction error $e[k]$ is then computed as $e[k] = x[k] - \hat{x}[k]$ and the estimated impulse response is updated by

$$\hat{h}[\theta_{k+1}] = \hat{h}[\theta_k] + \mu \cdot e[k] \cdot \frac{g[k]}{||g[k]||_2^2}, \tag{8.4}$$

where θ is the direction of the corresponding impulse response and μ is the step size, which is an important parameter for the NLMS HRIR estimation approach. The choice of μ can be considered as a trade-off between fast accommodation and noise-robustness. We refer to Haykin [66] and Enzner [46] for detailed information about the choice of the step size μ.

8.2. HRTF Measurement: Experiment

In this section, we compare the different methods of estimating HRIRs. To guarantee a fair comparison, a new evaluation method, called SNR_T is introduced beside the also conducted comparison by established evaluation criteria.

Experimental Setting

In the following, the experimental settings are described in detail. A KEMAR mannequin is placed on a turntable with a stepper motor in the institute's semi-anechoic chamber.

A loudspeaker in a distance of 2 m of the KEMAR plays the excitation signals which are recorded by the KEMAR's in-ear microphones. Table 8.1 gives an overview of the equipment we used in the experiments. The length of the HRIRs is 1024 samples and the sample rate is 48 kHz.

In the first experiment, the excitation signals for each method of HRIR estimation are recorded in discrete 5° azimuth-steps ("static"). In the second experiment a continuous azimuth-rotation ("dynamic") recording is conducted for the NLMS-type adaptive filter method. One rotation in the dynamic measurement approach lasts 98 s.

The estimated HRIRs of each method using static recordings of the first experiment are compared by different evaluation criteria. Finally, we compare the HRIRs generated by the dynamic recording with the spatially-discrete impulse responses by several instrumental evaluation criteria. Furthermore, the dynamic and the static measured responses are compared by a linstening test.

Objective Evaluation Methods

The fact that the different measurement approaches employ different kinds of excitation signals makes it difficult to find fair criteria to compare the HRIRs generated by the diverse measurement approaches. In addition to traditional ways of judging the quality of HRIR measurement environments (SNR_{Y_1}, SNR_{Y_2}) and measurement approaches (SNR_I, SNR_E), we propose a new evaluation method that is suitable to compare different HRIR estimation approaches, named "Test of HRIR Signal to Noise Ratio" (SNR_T).

An important issue for HRIR measurement is the laboratory environment. In the following, two characteristic values, namely, SNR_{Y_1} and SNR_{Y_2} are given that describe the institute's laboratory environment for HRIR measurement. The signal-to-noise ratio of the recordings is labelled SNR_{Y_1} and SNR_{Y_2} in this work. It is computed by

$$SNR_{Y_1} = 10\log_{10}\left(\frac{P_{signal}}{P_{noise}}\right), \tag{8.5}$$

where P_{signal} is the power of the recorded excitation signal and P_{noise} is the noise power of the recording system. All excitation signals were normalized to the same P_{signal}. SNR_{Y_2} is computed by

$$SNR_{Y_2} = 10\log_{10}\left(\frac{P_{HRIR_t}}{P_{HRIR_l} - P_{HRIR_t}}\right), \tag{8.6}$$

where P_{HRIR_l} is the power of a long HRIR (in our case 6000 samples), and P_{HRIR_t} is the power of the target sized HRIR (1024 samples).

SNR_{Y_1} is a descriptor of ambient and sensory noises in the lab environment, whereas SNR_{Y_2} describes the effect of reverberation in the lab room. Our semi-anechoic chamber [145] does not inhibit all low frequency reflections, which explains the relevance of SNR_{Y_2}.

Besides the SNR_{Y_1} and SNR_{Y_2} values which describe the laboratory environment, the performance of the different HRIR measurement approaches is compared by SNR_I, SNR_E and SNR_T.

SNR_I compares the power of the first samples of the HRIR to its maximum amplitude. The first samples of the estimated impulse response are expected to be zero, caused by delay due to sound propagation from loudspeaker to microphone. All deviations from zero in the first samples are therefore considered as noise of the HRIR measurement. SNR_I can be calculated by

$$SNR_I = 10\log_{10}\left(\frac{\frac{1}{N-K}\sum_{i=K+1}^{N} h[i]^2}{\frac{1}{K}\sum_{i=1}^{K} h[i]^2}\right) \tag{8.7}$$

where K describes the number of samples before the first wavefront hits the microphone and N describes the total length of the estimated HRIR. The SNR_I is also used as an algorithmic quantity in the context of step-size control of NLMS-type adaptive filters [66].

Our goal is to improve sound synthesis by an individual set of impulse responses. Therefore, the main purpose of our HRIR measurement lies in the rendering of an input sound signal s to a position in 3D space by convolving s with the corresponding set of HRIRs and playing back the resulting signal x' via headphones to the user. Beside the recorded excitation signals x_e from the different HRIR measurement approaches, we record further test signals x_t with the KEMAR's microphones. The recordings x_e and x_t include the KEMAR's real physical transfer function and serve as a reference for the virtually synthesized 3D sound x'. The SNR_E that compares the HRIR-convolved signal with the recorded excitation signal x_e is given by

$$SNR_E = 10\log_{10}\left(\frac{\sum_i x_e[i]^2}{\sum_i (x'[i] - x_e[i])^2}\right). \tag{8.8}$$

Enzner [46] uses SNR_E to evaluate the quality of dynamically measured HRIRs, due to direct relationship of SNR_E with the prediction error in the NLMS algorithm. One disadvantage of SNR_E is that it uses the method's excitation signal itself to compute the SNR_E, which makes it difficult to fairly compare different HRIR estimation methods. To achieve a fair comparison, SNR_T, calculated by

$$SNR_T = 10\log_{10}\left(\frac{\sum_i x_t[i]^2}{\sum_i (x'[i] - x_t[i])^2}\right) \tag{8.9}$$

uses independently recorded test signals x_t. White Gaussian noise serves as the HRIR input signal.

Objective Evaluation Results

The lab environment is characterized by SNR_{Y_1} and SNR_{Y_2}. The SNR_{Y_1} values for the horizontal plane are given in Table 8.2. It can be observed that the static methods have better SNR_{Y_1} values due to the fact that the turntable's stepper motor causes noise in the dynamic measurements. The mean SNR_{Y_2} is 37.1 dB for an HRIR-length of 1024 samples.

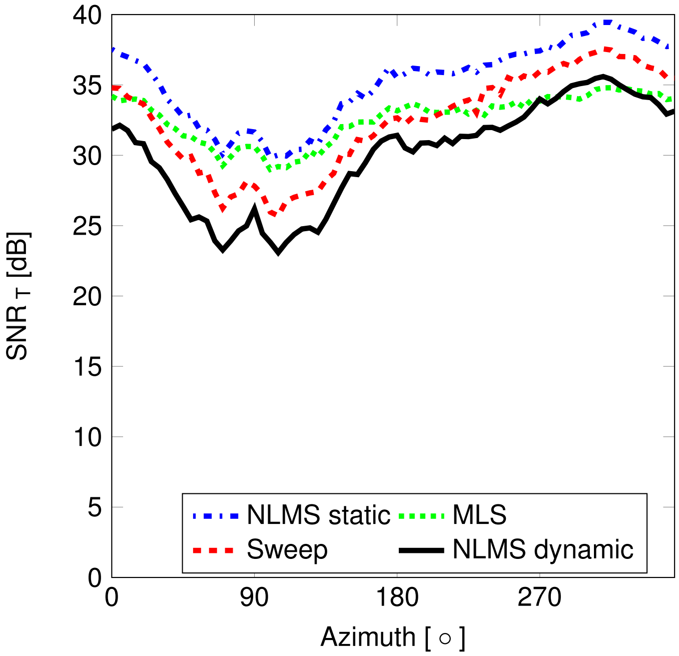

Figure 8.1.: Left ear SNR_T values for the different HRIR estimation approaches in the horizontal plane

	SNR_{Y_1} [dB]	SNR_E [dB]	SNR_T [dB]	SNR_I [dB]
Static NLMS	42.50	**35.58**	**35.20**	**50.83**
MLS	**43.81**	32.87	32.63	43.42
Sweep	41.51	27.23	32.42	40.43
Dynamic NLMS	31.26	28.49	30.01	35.06

Table 8.2.: Mean SNR values of the different HRIR estimation methods.

The unique feature of the adaptive filter method among the HRIR estimation methods is the ability to handle dynamic recordings with movements of the KEMAR during the recording procedure. For fair comparison of static and dynamic methods, two problems have to be tackled. The first problem is connected to the ability to precisely determine the position of the HRIRs. In the static case, our turntable is moved to a certain azimuth angle by a stepper motor with a built-in rotary encoder and the corresponding HRIR can be assigned to that azimuth position. Using the dynamic recording method, one has to compute the actual position of the determined HRIR using the turning speed of the turntable and the elapsed turning time. The second problem regarding the comparison of the static and the dynamic methods is the influence of the step size, affecting convergence speed and noise rejection rate. We solve these problems by choosing the dynamically generated impulse responses with the highest SNR_T for every azimuth position to ensure the best possible selection. This way, we can exactly compare the HRIRs generated by dynamic and static measurement methods. Figure 8.1 illustrates the SNR_T values in the horizontal plane. Using the MLS and exponential sine sweep technique signal to noise ratios (SNR) between $26\,\mathrm{dB}$ and $37\,\mathrm{dB}$ are reached. The SNR values are slightly lower for the azimuth angles between 0° and 180°, which can be explained by the fact that this is the contralateral side and consequently the excitation signal energy is lower than the excitation signal energy at the ipsilateral side. Beside MLS and exponential sine sweeps, static measurements were conducted using the NLMS adaptive filter method. The SNR_T values for the NLMS method are slightly better than the MLS method. Using the dynamic measurement method, the SNR_T values are slightly worse, especially at the contralateral side. This is due to the fact, that the turntable's stepper motor causes noise during the dynamic measurement while turning. The ratio of the noise at the contralateral side is greater due to the shadowing effects of the KEMAR's head, consequently reducing the power of the excitation signal. At the ipsilateral side, the ratio between excitation signal and the turntable's noise is higher resulting in better SNR_T values.

Listening Tests

As illustrated in Figure 8.1, the SNR_T values for the static estimation approaches are higher than the dynamic ones. We aim at acoustic teleconferencing applications, therefore, the most important question is whether subjects can distinguish between test signals filtered with different sets of HRTFs generated by the afore mentioned dynamic and static transfer function estimation approaches in listening tests. To answer this question, we apply an ABX double blind test [18].

Four sets of HRIRs, generated by static MLS, static sweep, static NLMS and dynamic NLMS, respectively, are used to virtually synthesize a noise burst sequence to move around a listener in the elevation zero plane.

According to the test guidelines in [18, 75], a fifth test signal is generated and serves as a hidden reference in order to ensure if the test subjects are watchful along the test procedure. Each static test signal was compared 16 times to the dynamic ones played back

	MLS	Static NLMS	Sweep	Dummy
correct answers	0.48	0.47	0.49	0.95

Table 8.3.: ABX test results: The test listeners could not hear differences between the test signals that are synthesized with the differently measured HRIR databases.

to the 21 listeners with *Beyerdynamic DT 990 Pro* headphones in our audio laboratory. The test subjects were introduced to the test setting according to the test guidelines. In each test round, the listener has to decide whether the presented signal *X* is identical to a presented signal *A* or signal *B*. In each round, *X* is actually a copy of randomly played back *A* or *B*. According to Table 8.3, no one of the test listeners could reliably hear any difference between the dynamic NLMS, static MLS, static sweep and static NLMS generated 3D test signals (with a confidence of 95 %). Three participants were excluded due to the fact that the hidden reference was not assigned properly.

8.3. Concluding Remarks: HRTF Measurement

In this chapter, I address the problem of acoustic HRIR measurement. We compare three static HRIR measurement methods, namely exponential sine sweep, MLS and the static NLMS method with the dynamic NLMS method by different objective and subjective evaluation criteria. The different SNR values of the dynamic NLMS method are slightly worse than the SNR values of the static methods. However, listeners could not differentiate between the differently measured HRIR databases in listening tests. Since a discrete HRIR measurement with 1° resolution in the horizontal plane typically lasts about 50 min in our lab, whereas a dynamic NLMS-based measurement with quasi-infinite resolution takes only 2 min, the dynamic measurement is a promising new alternative to static approaches. Therefore, I decide to use the dynamic measurement approach to set up a LDV HRTF database for further investigations on immersive playback of teleconference speech contributions.

Part III.

System Evaluation

To finalize the conference system modeling I quantify the benefit of the speaker assignment and immersive playback efforts that were made to enable the remote listener's virtual audio participation of conferences.

The individual components of the whole teleconferencing system are evaluated by the typical characteristics of the subproblems to be solved to achieve an immersive teleconferencing system. In Part I of the thesis, the problem of channel assignment is solved by sound source localization, separation and speaker recognition. The possible sound source localization approaches for the developed microphone array prototypes are judged by experiments that determine the localization success rate and the mean angular error of the localization approaches. The sound source separation algorithm's signal to distortion, signal to interference and signal to artifacts ratios then assess the separation performance. The diarization error rate finally judge the speaker recognition and the developed channel assignment approaches. Part II of the thesis presents various methods to achieve immersive playback at the remote site of the conference with respect to the listener's individual HRTFs. The quality of a HRTF database for virtual playback is often identified by listening experiments that determine the localization error, e.g., to choose among two HRTF selection methods in Chapter 6. In the field of generating individualized HRTFs by regression methods as described in Chapter 7, the spectral distortion is a frequently used measure to compare computed HRTFs with the ground truth. In order to compare various approaches of acoustic measurements of HRTFs, we successfully developed different signal to noise ratios [141] that are presented in Chapter 8.

Each of the respective evaluation methods of the teleconferencing system's subproblems has its justification to describe the quality of solving the particular subproblems. However, the individual key figures do not adequately describe the whole system. Therefore, the final part of my thesis deals with the evaluation of the speaker assignment algorithm in combination with different methods of immersive playback to answer the following questions:

- How do remote teleconference participants judge the possibility of spatialized playback of conference contributions?
- Does the speaker assignment algorithm introduce annoying artifacts that render the achieved immersive playback useless?
- Do the channel assignment and immersive playback improve the perceived quality of experience (QoE)?
- Do the channel assignment and immersive playback improve the efficiency of teleconferencing?
- Is the effort of measuring individual HRTFs justified for the teleconference scenario compared to HRTF selection, regression or standard KEMAR HRTFs?
- Is it worth to further investigate high accuracy head-tracking for teleconferencing?

- How does the introduced channel assignment algorithm perform compared to perfectly separated conference contributions?

I decide to divide the evaluation of the developed system into two main criteria, namely, the quality of experience (QoE) and the cognitive load (CL) for the teleconference system. Besides QoE and CL, a listening test is made to judge the subject's localization performance of the different HRIR databases.

9. System Evaluation Concepts

In this chapter, three concepts of evaluating the speaker assignment and immersive playback are introduced, namely, the quality of experience concept, the cognitive load concept and the sound localization performance.

9.1. Sound Localization Performance

A frequently used method to judge the performance of a HRTF dataset for synthesizing sound sources is to conduct subjective sound localization (SL) listening tests. Therefore, I decide to also evaluate the SL performance of probands with the different HRTF datasets that are considered to be deployed in the developed teleconference system. There are different methods seeking to reliably measure the SL performance of listeners in hearing tests.

One method is the so-called identification method, described in [112, 113]. The possible sound sources are represented by real speakers at predefined positions in the test environment. Then the test subjects listen to the playback of the loudspeakers as well as to the virtually synthesized sound image and mark the supposed sound source position on a form where the predefined positions are documented. Thus, the method is intuitive and there are no confusions due to proprioception of the listeners. However, the proband rather has to decide among the predefined sound source positions than to actually localize the virtual sound source.

Another method to conduct SL listening tests is to ask the probands to tell the perceived azimuth and elevation angles of a virtual sound image to the experimenter or to mark it on a coordinate system on a questionnaire. Therefore, the probands have to localize the sound source and then map the perceived direction to coordinates of the written or digital form [13, 34]. The drawback of this method is that listeners have individual scales of transforming the perceived direction information to azimuth and elevation coordinates unless the test listeners are trained to correctly map the perceived direction to azimuth and elevation coordinates.

The direction of a perceived sound source can also be detected by tracking methods such as head tracking and eye tracking. The eye tracking method seeks to measure the saccadic eye movements that are supposed to hint towards the perceived direction of the sound source. The reliable detection of the eye movements can be done with optical tracking or, if a shaded test environment is required, with magnetic detection of a scleral search coil in the proband's eye [52, 68]. Besides the elaborate of the eye tracking, the listener's head has to be fixed during the tests to exclude false directions caused by head movements. Due to technical reasons, only directions in the front hemisphere can be evaluated.

The head tracking method [107] to detect the direction of a sound source in listening tests is similar to the idea of the eye tracking method. In the head tracking method, the orientation of the proband's head is detected which should indicate the perceived direction of the sound image. However, the orientation of the head does not necessarily correlate to the perceived direction of a sound source, e.g., when the proband does not look straight ahead.

To overcome the problem of reliably capturing the probands direction impression, the optical pointer method is invented where the test subject uses a device, e.g., a track ball to direct a laser pointer that is detected with preinstalled sensors on a discrete grid to unveil the position of the virtual sound source [98, 153]. The indirect control of the laser pointer by an input device decouples the localization task from the human's motor function to avoid proprioceptive effects that could disturb the localization capturing task. The optical pointer method only allows direct testing of virtual sound sources at predefined locations in the frontal hemisphere due to the experimental design.

A further development of the optical pointer method is the laser pointing method. Contrary to the optical pointer method, the proband intuitively points at the supposed sound source and gets optical feedback by a laser dot. The direction of the gesture then can be determined by tracking markers fixed at the laser pointer [122]. This way, all azimuth and elevation angles at a constant radius can be chosen by the proband. One drawback of this method could be that the listener's movement influences the direction choice.

The SL performance gives feedback of the ability of the proposed HRTF datasets to virtually add direction information and the SL listening experiment only covers the playback aspect of the developed immersive 3D teleconferencing system without regarding the data acquisition aspect. Moreover, it is questionable, if the pure localization error consideration actually reflects a fair evaluation of a teleconferencing system and thus, further evaluation concepts need to discussed to evaluate the teleconferencing system.

9.2. Quality of Experience

The perceived quality of a teleconference system user is one of the most important aspects to judge the performance of my ideas to achieve an immersive 3D teleconference for a remote conference participant. In contrast to objective evaluation criteria like the localization success rate for the sound source localization, the SIR, SAR and SDR values for the sound source separation, or the DER for the speaker recognition, the measurement of the perceived user quality of experience QoE is not straightforward.

First of all, the expression "quality" can be interpreted differently. Pioneering attempts to define "quality" were made in chemometrics and are overviewed by [108], where four different aspects of quality are considered:

- **Qualitas**: Quality due to objective characteristic properties, e.g., color or material of a device.
- **Excellence**: Quality due to subjective intuitive judgement of a device.
- **Standards**: Quality defined by the degree of achieving predefined requirements of a device.
- **Quality as an Event**: Quality defined by the subjective impression by using a device.

The first two quality aspects are either objective or subjective criteria. The third definition seeks to combine the first two aspects by finding criteria that reflect the subjective judgment of characteristic properties of using a device. The definition of quality as an event describes the usage of a device with a certain quality by a subject that feels a certain excellence of usage.

Concerning the teleconference application, the fourth definition "Quality as an Event" provides a good concept that includes all components of the developed conferencing system that finally lead to the chance of immersive playback of the conference contributions. The concept further can be specified with respect to sound quality. According to [97], sound quality can be assessed by presenting auditory images to subjects that judge the sound samples by their degree of satisfaction, compared to other presented auditory images. Furthermore, it is suggested to evaluate the sound quality by parametrizing the auditory image in order to take the multidimensional character of sound events into account [15, 147].

However, evaluating different parameters of an auditory image could result in discrepancies between an expert listening test designer and naive probands. Therefore, it could be beneficial to elaborate a test vocabulary together with the test subjects, which additionally causes huge efforts in advance of the actual listening test. For further information regarding trainings in advance of the listening tests refer to [101].

In contrast to the direct questioning of test subjects, a further approach is to conduct psychoacoustic experiments in order to obtain characteristics that are connected to physical measurements [49]. Another approach to evaluate the HRTF-based sound synthesis of the teleconferencing system is to evaluate the quality of the HRTFs by the key property of the HRTFs, the direction dependency of the transfer functions. Therefore, the quality determination of the HRTFs is often based on listening tests that seek to evaluate the localization ability of the probands with the respective HRTF datasets [13]. Section 9.1 gives a detailed overview about this issue. Moreover, speech comprehensibility [24, 41], speaker identification [17] or task performance [83] can be a measure of quality for a teleconferencing system.

The most utilized concept to evaluate an acoustic teleconference system seeks to quantify the quality by the degree of achieving predefined standards. This concept is referred to as quality of service (QoS). In [76] the ITU defined criteria that are mainly related to the performance of a system, e.g. packet loss or mouth to ear delay. Furthermore, a concept

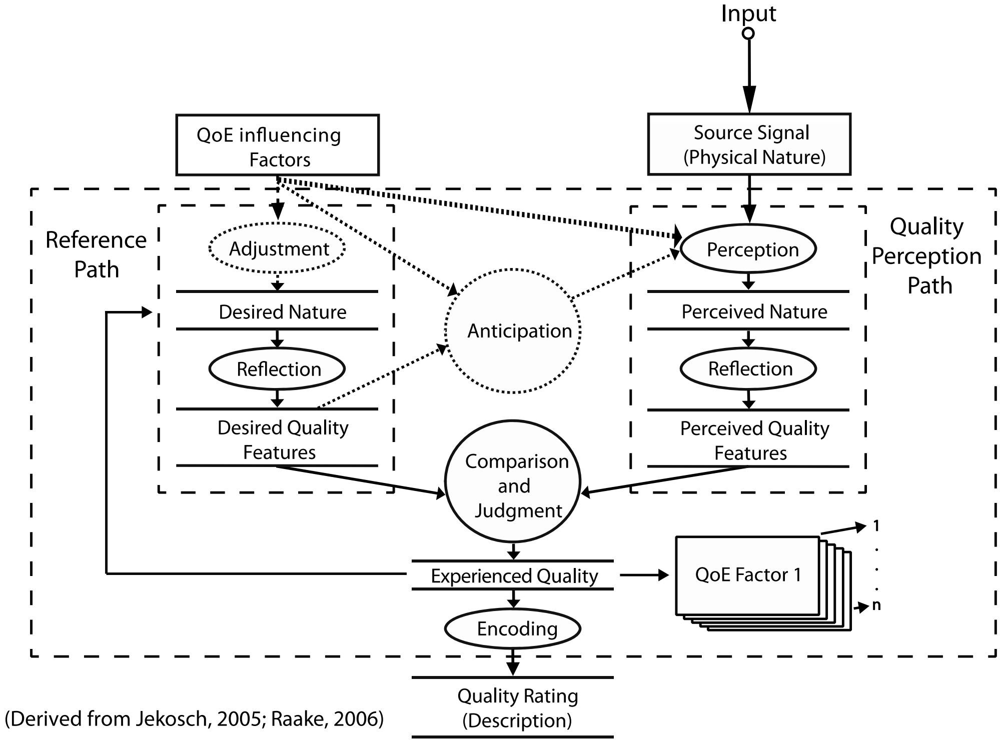

Figure 9.1.: Schematic overview about the proband's QoE decision-making process [28]

.

of quality of service experienced by customer (QoSE) is mentioned in [76] that incorporates the subjective user opinion that is connected to criteria such as the speech quality [78] or the mean opinion score (MOS) according to [80]. The idea of QoSE to evaluate our system is connected to the concept of evaluating the system by its QoE.

The QoE concept of evaluating our system is related to the idea of considering quality as an event [108]. In [28], it is attempted to define the QoE by describing the meaning of experience and quality in the context of evaluating a system. In contrast to the QoS, important factors are the subjective perception and judgement of a user and the comparison of the quality by references. The QoE decision-making process is shown in Figure 9.1. During the QoE evaluation, the probands judge the perceived quality by an individual, internal parametrization process, the QoE factors illustrated in Figure 9.1, which rather correspond to a customer's quality judgement than an evaluation of different parameters by expert listeners. In accordance with [28], the QoE is defined as

> „ *(...) the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect*

to the utility and/or enjoyment of the application or service in the light of the users personality and current state.“

9.3. Cognitive Load

Besides the afore discussed SL and QoE concepts, another idea of evaluating our teleconferencing system is the so-called usability of the system. The usability describes the degree to which the use of a device fulfills predefined requirements regarding satisfaction, effectiveness and efficiency [115].

The usability concept has an overlap with the QoE concept, especially in the satisfaction dimension. However, the QoE is the result of a judgement process, without taking subconscious processes into account, e.g., the effectiveness. An example for the effectiveness in the context of a teleconferencing system is the amount of content that can be remembered by the participants after the conference. The efficiency of a teleconference system describes the resources that have to be invested to reach a certain goal, e.g., the amount of concentration to follow the conference. The terms effectiveness and efficiency can be summarized to the terms mental workload or cognitive load (CL). There are two different disciplines to analyze mental workload: the ergonomic analysis and the cognitive load theory.

The ergonomic analysis considers teleconferencing as work activity, therefore, a teleconference system can be regarded as a tool. The ergonomic analysis seeks to design tasks, as far as technically and organizationally possible, with respect to physical and mental workload considerations [167]. Besides usability, an ergonomic teleconference system should also keep the mental burden as low as possible. If the physical and mental burden during performing a task is too high, different compensation activities are activated, e.g., exhaustion, monotony, stress and mental satiation [167]. All mentioned compensation activities are thinkable during a teleconference, but especially the mental exhaustion is a factor that can be tackled with advanced teleconferencing technology such as immersive playback of the conference participants.

According to [156], the CL theory describes a model of human information processing and learning theory and how cognitive capabilities enable or restrict learning processes. One key assumption is the restricted capacity of the human's working memory. The CL-theory then derives recommendations from the restricted working memory capacity to design learning situations that do not overload the working memory. The working memory can further be divided into three parts [10]: the attentional control, the buffer memory for speech and acoustical information and the buffer memory for visual and spatial information. According to [11] the separated individual buffer memory sections for speech and spatial information can be a explanation for the advantages of binaural sound synthesis, e.g., in teleconferencing systems.

In order to measure the mental workload, the measurement process can be divided into three categories:

- **Somatic**: Measuring the physiological reaction of the subjects body.
- **Subjective estimation**: Measuring the perceived mental workload.
- **Performance**: Measuring the task performance of the subjects.

The physiological reaction in an experiment, e.g., a teleconference scenario, to evaluate the mental workload can include, the measurement of the heart rate and respiration rate, the blood pressure and electrical resistance of the skin [160, 167]. One advantage to use the physiological reaction to determine the mental workload is that a prior training of the test subjects is not required. Furthermore, the measurement process can deliver continuous and objective data during the experiment. However, according to [105], measuring the physiological reactions of a subject can hardly be applied for hearing tests. Moreover, the interpretation of measured physiological reactions is very difficult and the results can vary due to many different reasons such as the proband's age, health condition or coffee consumption.

Due to the fact that probands usually are aware of the mental workload, one of the most commonly used methods to determine the mental load of a subject in a certain task is to use questioning methods. Therefore, there are several well established guidelines to design a questionnaire to determine the mental load. The ITU developed recommendations to conduct subjective quality evaluation tests, including a discrete listening effort scale [80]. In [157], evaluations about hearing constraints in 3D teleconferencing were conducted, where the scale of the questionnaire are extended at the scale's edge [115] to avoid contraction bias [187]. Another possibility is to use the NASA Task Load Index (NASA-TLX) [63, 64], which seeks to determine different dimensions of effort. The different dimensions are then merged by an individual weighting. Therefore, an additional effort is necessary after the actual questioning to determine the weighting. Moreover, the probands have to be familiar with the test scale, which requires extra training in advance of the questioning.

The third category of measuring mental workload, is to measure the task performance of the probands in a teleconferencing situation. The task performance can be divided into a primary task, e.g., remembering the content of a conference, and into a secondary task, e.g., the multitasking ability in a conference. In this work, I want to focus on the hearing process within a teleconference situation rather than on the conversation aspect. Identified primary tasks in the teleconference scenario are the speech comprehensibility and the speech comprehension.

To determine the speech comprehensibility, isolated words [23] or sentences [41] are played back to a listener in order to measure the speech reception threshold. In the military sector, another method was developed to measure the speech comprehensibility, called coordinate response measure [19], where the proband has to detect an individual call sign in an audio stream. The comprehensibility depends on the response time and correct answer of asked content-questions on the played back audio signal.

The speech comprehension describes the amount of conversation information that can be memorized by a conference participant, including the conference content and the conferee who contributed the respective conference content [115]. Speech comprehension evaluations were already conducted in 3D teleconferencing systems. The probands in such tests had to remember which conference participant contributed a certain information and should write down the respective viewpoints of the conference participants [87, 157]. Testing the memorized contents of a teleconference is a useful objective measure, however, it is found that the standalone measurement of the recall performance is unsatisfactory due to the fact that stressful situation can be compensated with higher efforts [27].

In order to detect these compensation efforts, one can introduce a secondary task to describe the mental workload of a teleconference user. The secondary task is an additional artificial task that has to be processed simultaneously to the primary task to tackle the problem of temporary higher compensation efforts of the test subject.

10. Listening Experiments

In this chapter, the complete system is evaluated extensively by the afore mentioned criteria, namely, the quality of experience (QoE), the cognitive Load (CL) and the sound localization (SL) performance of the test subjects.

10.1. Overall Experimental Settings

This section gives an overview of the Institute for Data Processing (LDV) audio laboratory and the LDV HRTF database which are used in the SL, the QoE and the CL listening experiments.

LDV Audio Laboratory

In order to quantify the benefit of the immersive, HRTF-based playback approaches of the assigned audio channels within a teleconference scenario, I have constructed an audio laboratory at the Institute for Data Processing (LDV) that enables us to measure individual head-related impulse responses. Construction considerations to achieve a good compromise between costs, required space and anechoic conditions are described in detail in [145].

For the listening experiments, we recorded the LDV-HRIR databases of 35 human subjects in our laboratory environment. Each person's database was recorded for six different elevation planes with an azimuth resolution of 1°.

Figure 10.1 shows the semi-anechoic chamber at our institute. The purpose of the chamber is to terminate acoustic reflexions and diffractions in the room as well as noise around our institute to evaluate algorithms and techniques for channel assignment. Next, the lab is equipped and used for the measurement of HRTFs. Moreover, listening experiments are conducted in the semi-anechoic chamber to guarantee identical acoustic conditions for the test subjects.

The reverberation time of the LDV audio lab is $t_{60} = 0.08\,\mathrm{s}$ and the lower cut-off frequency is 160 Hz within a distance of 1.3 m from the laboratory's center. Due to the directional characteristics of our sound source, only frequencies from 80 Hz...200 Hz could be considered in accordance with ISO 3745. Analogous to [53, 184, 185] we expect higher frequencies to meet the ISO 3745 norm, too.

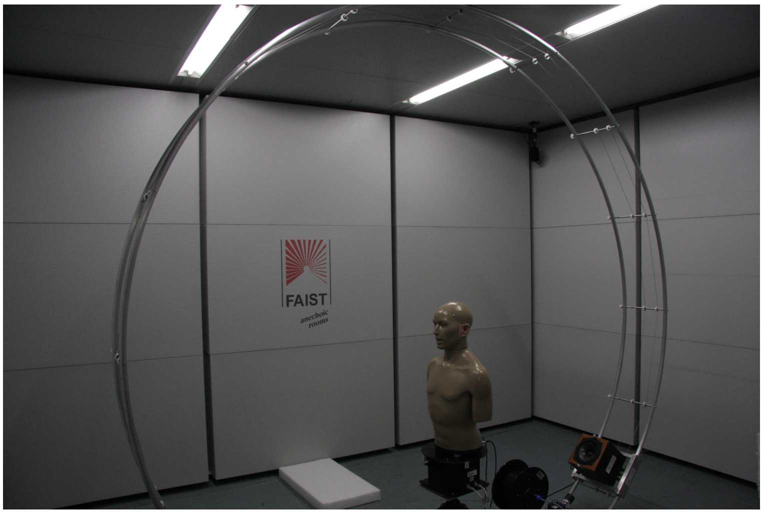

Figure 10.1.: Audiolab at the Institute for Data Processing

HRTF Measurement of a Listener

In Chapter 8 several HRTF estimation approaches are analyzed. To time-efficiently obtain an individual dense HRTF dataset for the users of the developed conferencing system, I suggest the dynamic approach using NLMS type adaptive filters. In order to generate individual sets of HRIRs with the dynamic approach, several tasks have to be accomplished.

Listener Measurement Position

To guarantee reproducible and interpersonal comparable measurements, the listeners have to be fixed to a certain position. The measurement positioning has to be achieved with slender devices that do not disturb the sound field and consequently degrade the HRTFs of the listener. In literature, there are a number of fixation suggestions, e.g.:

- The listener has to stand upright on a turntable during the measurement, fixed by a back rest [116].
- The listener sits on the floor and the loudspeaker moves around the listener [162].
- The listener is fixed to a turnable chair [2, 96].

Any movement of the listener's head changes the direction dependent reflexions and diffractions off the head and torso and consequently take effect on the measured head-related impulse response [130]. To overcome the problem of head movements during measurement, there are approaches to transfer the responsibility for the head positioning to the listener by providing a camera, a screen and an automatic release for the test subject

[96] or to attach markers to the test subject in combination with a control screen [116]. In [96, 121] the influence of the back rest and head rest is observed and reflexions of the measurement construction are observable but accepted.

Microphone Measurement Position

Besides the test subject's position, the position of the microphone during the measurements plays an important role. The most prominent microphone positions for HRTF measurement are in the ear canal, at the entrance of the ear canal and at the blocked entrance of the ear canal [61, 67]. The blocked ear canal method for individual HRTF measurement is attractive, because the direction dependency of the transfer functions is included [3, 61, 72] and the fixation of the microphones is easier to arrange and reproducible without the risk of harming the ear drum. Furthermore, the response of the ear canal is listener dependent, but not direction dependent [61, 67].

In [129] the positions of the microphones are studied. It is found, that transfer functions, measured at neighboring microphone positions in 2 mm distance, show considerable transfer function differences. Besides the variation in positioning the microphones, there also exists a variety of utilized microphone types. *Sennheiser KE 4-211-2* microphones are used in [116], *Bruel and Kjaer 4182* probe microphones in [116], *Etymotic ER-7C* probe microphones in [3] and *Knowles FG3329* are utilized for the IRCAM database [72].

LDV-Database Measurement Conditions

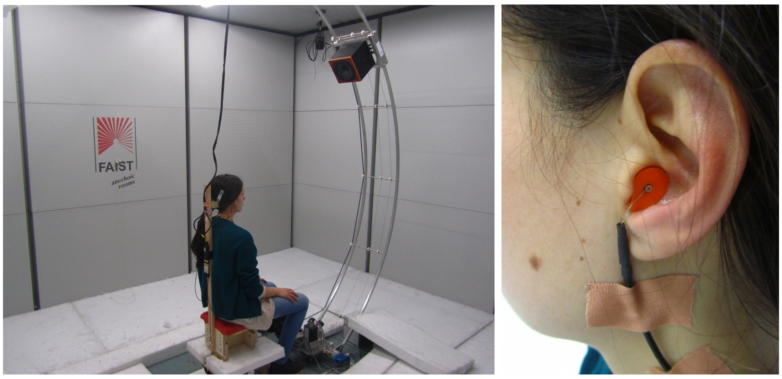

Figure 10.2.: The Figure shows the measurement position in the LDV lab (left) and the ear plugs with fixed microphones (right)

For the subjects' acoustic measurement position, we decided to construct a turntable with a seat and a thin back and neck rest that enables the listener to hold still in an unstressed position during the measurement process which is illustrated in Figure 10.2. The sitting position enables us to cover a variety of elevation positions (-50° to 230°) at constant speaker distance (130 cm) in our lab. Furthermore, the sitting position during the measurement is the most likely teleconference user position. Besides technical aspects, we think that we reach the most immersive impression in a teleconference scenario by choosing the described measurement position for the listener. The constructed turntable with a stepper motor allows for continuous azimuth rotation, appropriate for the chosen dynamic NLMS HRTF measurement approach. To automatically adjust the elevation positions of the speaker, we constructed a metallic arch with a diameter of three meters, where the loudspeaker is placed on a movable carriage. The carriage with the speaker can remotely and automatically be moved with the arch's stepper motor.

To ensure reproducibility and interpersonal comparability to a maximum possible extent, we decided to block the ear canal with silicone plugs, illustrated in Figure 10.2, that are custom made for each listener as suggested in [72]. As it has similarly been performed in [72], we choose the *Knowles FG3329* microphones because of the space-saving design and the applicability in combination with the silicone ear plugs.

At the end of each measurement session, an extra measurement was conducted to obtain the compensation impulse response (cIR). For this purpose, the microphones were fixed at the center of the hoop and the impulse response was measured by the MLS method. To overcome inversion instabilities of the cIRs, a frequency-dependent regularization was applied, where only small filter coefficients are regularized while stable filter coefficients remain unaffected [35]. A detailed description of the LDV Database can be found in [144].

10.2. Sound Localization Evaluation

The first part of the evaluation campaign consists of SL listening tests to determine the ability of the different datasets to virtually synthesize sound sources at certain directions around the listeners.

Experimental Settings: SL

In Section 9.1 different methods to experimentally determine the SL performance in listening tests are introduced. In this work, the listening test is conducted in accordance with the laser pointing method. The laser pointing method is characterized by a self describing, intuitive handling without the need for the proband to project the listening results to different coordinate systems on a evaluation sheet. Furthermore, the method allows for automatic testing of the front and back hemisphere.

Laser Pointing Method Setup

Different stimuli are virtually synthesized by a convolution engine that is fed with the different HRTF datasets. The stimuli then are played back to the listeners with headphones. A laptop computer serves as an evaluation interface that collects the directional estimations of the probands which are captured by pointing a wireless presenter towards an estimated sound direction and confirming the choice by pushing the wireless presenter's buttons. The presenter is equipped with a laser pointer that is reflected by a fleece that forms a cylinder with a diameter of 2.5 m around the listener. Tracking markers allow us to compute the azimuth and elevation information that is automatically captured by the evaluation interface. Figure 10.3 shows the test setup and a proband doing the listening test.

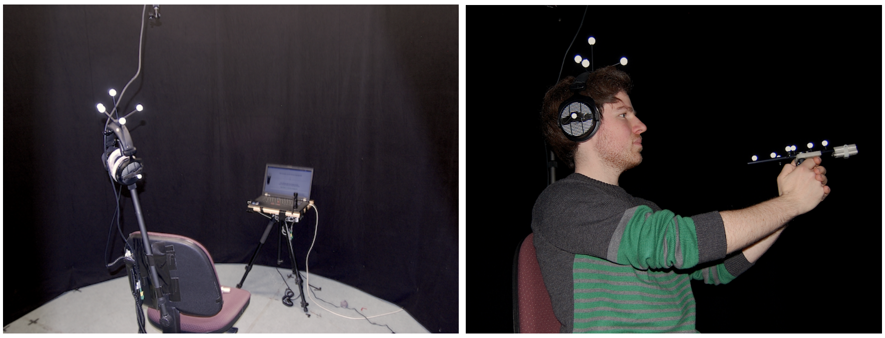

Figure 10.3.: The Figure shows the SL experimental setup in the LDV lab (left) and a proband during the SL listening tests (right)

Playback System Variations

There are different components of the convolution engine at the remote listeners site that can be varied:

- **High Quality Head Tracking (HQHT)**:

  To enable dynamic sound synthesis with respect to the listener's head movement, head tracking is required. In the HQHT-scenario, the head rotation of the listener is captured by a professional tracking system built by *Advanced Realtime Tracking (ART)*. The audiolab ART-installation consists of three infrared cameras that enable us to track reflecting tracking bodies that are fixed at the headphones with 6 degrees of freedom at a frame rate of 30 Hz to precisely determine the head pose of the listener.

- **No Head Tracking (NHT)**:

  In order to quantify the benefit of head tracking in the teleconferencing scenario, I decide to also evaluate a teleconferencing situation with a static sound field that does not take the user's head movements into account.

- **Non-Individual HRTF Database (K-DB)**:

  In the K-DB case, the convolution engine is fed with the KEMAR HRTF dataset for each listener. The K-DB database is measured in the institute's audio laboratory. Instead of listeners, a widely used KEMAR mannequin by *G.R.A.S. Sound & Vibration* is used to generate a set of HRTFs. One major advantage of utilizing the K-DB for the immersive playback of a teleconferencing system is the simple use of just one dataset. Therefore, there is no further effort required by the listener to configure the playback system or to upload customized sets of HRTFs. The trade-off that comes with this usability advantage is the non-individual nature of the sound synthesis, which is often considered as the major drawback of the K-DB.

- **Selection HRTF Database (S-DB)**:

  In the S-DB case, an individually chosen HRTF database is used for sound synthesis. The selection process, described in Chapter 6, consists of a swiss style tournament listening test to select the most appropriate HRTF dataset for each listener in accordance with [81]. For fair comparison, the listeners have to choose among twelve subject's HRTF datasets of the LDV-DB that serve as selection pool. The subjects that are in the selection pool are excluded from further experiments. The advantage of the S-DB is that one can offer customized HRTF-based sound rendering without acoustical measurements.

- **Regression-Generated Databases (R-DB)**:

  A customized set of HRTFs can be generated by measuring certain anthropometric data of the listener which are used to compute an individual HRTF dataset for the conferee. In this work, listening tests with three different regression options are conducted, namely, the 2DPCA, the PLSR and the GLRAM, specified in Chapter 7. For each proband, a customized HRTF dataset is computed using of the anthropometric data and the LDV-DB that serves as a training set for the regression. Of course, the proband's actual measured dataset is not part of the training set.

- **Individually Measured HRTF Database (I-DB)**:

  In the I-DB case, the convolution engine is fed with the acoustically measured HRTFs of the proband.

Experimental Design

The sound localization listening experiments consist of two parts, session L1 and session L2, which are conducted in separate evaluation sessions. The two sessions are performed by a different group of probands in order to avoid learning effects. In L1 the test listeners determine the perceived direction of the incoming sound with the tracking options HQHT and NHT which are combined with the K-DB, S-DB and I-DB. On the basis of the L1 results, the probands are confronted with one stimulus in session L2 and the two head

Test sound	Stimulus 1						Stimulus 2					
Head tracking	HQHT			NHT			HQHT			NHT		
HRTF-DB	K	S	I	K	S	I	K	S	I	K	S	I

Table 10.1.: SL(L1): Schematic overview about the different stimulus treatments for listening test part 1.

Head tracking	HQHT		NHT	
HRTF-DB	K	R (2DPCA, PLSR, GLRAM)	K	R(2DPCA, PLSR, GLRAM)

Table 10.2.: SL(L2): Schematic overview about the different stimulus treatments for listening test part 2.

tracking options in combination with the afore mentioned R-DBs (2DPCA, GLRAM, PLSR) and the K-DB that serves as a common reference in L1 and L2.

Stimulus Treatments

Table 10.1 illustrates the twelve possible stimulus treatments for L1 and the eight stimulus treatments for L2, which are tested in a full factorial, within-subject design. Full factorial experiment design specifies a listening test, where each proband is confronted with all possible stimulus treatments and within-subject design describes that each proband listen to each variation within the stimulus treatments [12].

L1 confronts the listeners with different measured HRTF-datasets that are used to virtually synthesize the stimuli to different positions around the listener. In the L2-session, customized HRTF-datasets computed by regression methods are utilized to generate HRTF-based 3D sound. Both sessions include the K-DB that serves as a reference in order to enable us to draw conclusions based on the results of the two sessions.

Stimuli

In the sound localization listening experiments, two kinds of stimuli with a length of 2 s are used: A noise burst stimulus (stimulus 1) and a male speech stimulus (stimulus 2). Broadband-noise stimuli are often used for listening tests since they contain energy in every frequency band. Therefore, all possible frequency dependent cues of the human listening system can be addressed, e.g., high frequency dependencies of monaural cues [16].

We additionally apply a speech stimulus that was recorded in the LDV audiolab as an alternative to the noise bursts in order to also present a speech stimulus in the listening experiments that is representative for a teleconferencing situation.

Stimuli Presentation Sequence

The stimulus treatments illustrated in Table 10.1 can be segmented into four stimulus treatment groups: The HQHT-stimulus 1 group, the HQHT-stimulus 2 group, the NHT-stimulus 1 group and the NHT-stimulus 2 group. The probands are also divided into four groups and the different stimuli within a group are tested group by group meaning, that the different datasets and directions are tested with one head tracking option and one stimulus. Then, the datasets and directions are tested within the next group of tracking and stimulus option.

In order to avoid listening test results that are dependent on the stimuli presentation sequence, a balanced latin square design [20] of the groups is applied to determine the stimuli presentation sequence of the groups and of the subjacent HRTF datasets. Thus, each of the four groups of test subject is listening to a different stimuli sequence, moreover, it is ensured that succeeding stimuli treatments are never presented in the same order to different probands.

SL: Experimental Procedure

The SL listening experiment is conducted in the LDV audiolab [145]. The probands sit in front of an evaluation interface that we programmed for the SL evaluation campaign. The different stimuli are played back with *Beyerdynamic DT 990 Pro* headphones with attached tracking markers of the HQHT system. The proband navigates through the test procedure using a presenter with mounted tracking markers to determine the listeners choice of the perceived direction of the virtually synthesized sound source which is played back by a *Roland UA 25 EX* audio interface. An oral introduction of the evaluation supervisor and a demonstration scenario makes the listeners acquainted with the laser pointing method and the test procedure. For each stimulus treatment, 40 different angles around the listener are rendered, eight different azimuth angles and five elevation angles, arranged in accordance with [122]. Each proband has to pass the four different stimulus treatment groups in L1 with a listening pause of several hours after the first two groups. The determination of each listener's I-DB was done in a separate measurement meeting several weeks before the actual SL evaluation. The individual selection of the S-DB was also made several days in advance of the actual SL evaluation sessions.

In L2, another group of test subjects conducts the listening tests with the stimulus treatments illustrated in Table 10.2. The regression generated HRTF databases are computed before the listening tests by using a set of anthropometric parameters of the probands. Contrary to L1, only speech stimuli are used.

Experimental Results: SL

The SL listening experiments were conducted with 40 probands divided into two subgroups for L1 and L2, respectively.

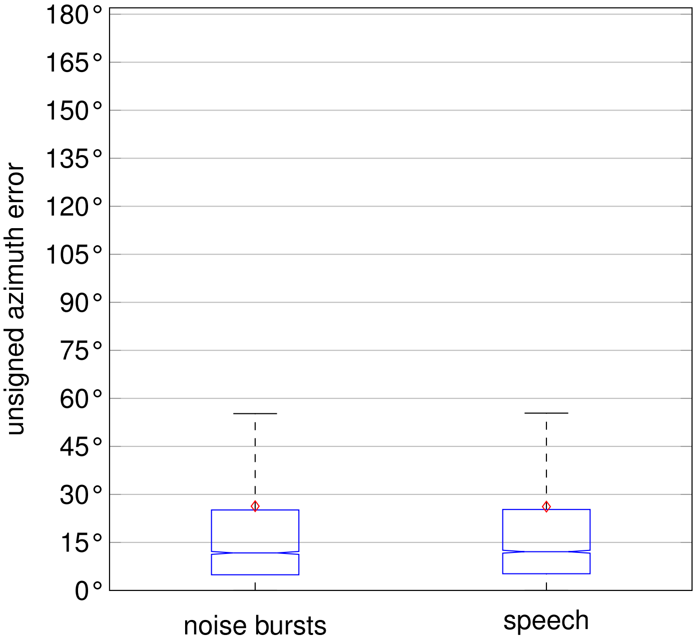

Figure 10.4.: SL (L1) evaluation of different stimuli

In the SL listening experiments the localization error is used to objectively quantify the performance of the different playback options of the developed teleconferencing system.

One way to illustrate the SL results is to use Boxplots and the statistical analysis by a so-called ANOVA (analysis of variance) suggested by [12]. In this work, we consider results with p-values of lower than $p = 0.05$ as statistically significant. An example for Boxplots can be seen in Figure 10.4. The horizontal blue line in the Boxplots denote the median of the listeners SL judgements, the box represents the 25th and 75th percentile, respectively. The box's notch represents the 95 % confidence interval and the judgements outside the whiskers are considered as outliers. The arithmetic mean of the judgements is illustrated by a red diamond shape.

Listening Test L1

In the first SL-evaluation campaign, the 20 probands are confronted with stimulus treatments summarized in Table 10.1.

SL(L1): Stimulus

The stimuli are compared as shown in Figure 10.4. The listening test results show that there are no statistically significant differences between the localization experiments using

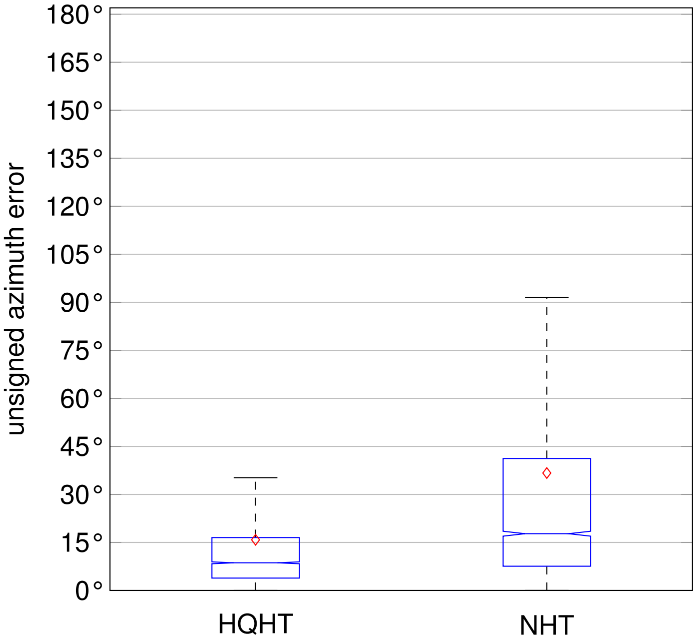

Figure 10.5.: SL (L1) evaluation of the different head tracking options HQHT and NHT

noise bursts and speech. The mean angular error for noise bursts and speech are 26.32° and 26.18°, respectively. The amount of front-back and back-front confusions in the noise burst scenario is 15.1 % compared to 22.8 % in the speech scenario.

SL(L1): Head Tracking

Figure 10.5 illustrates the difference in localization performance using HQHT or NHT. Obviously, the localization error is statistically significant lower ($p < 0.01$) if the proband's head movements are reliably captured and included in the sound synthesis, even if the stimulus length of 2 s does not allow the listener to strategically use movements to face the sound sources.

The mean angular error using HQHT is 15.82° and 36.68° using the NHT playback option. Furthermore, the amount of front-back and back-front confusions can be reduced from 20.8 % to 4.5 % in the HQHT option compared to the NHT option.

The results show that it is advantageous to enable head tracking in combination with the different HRTF databases that we generated by the afore mentioned approaches.

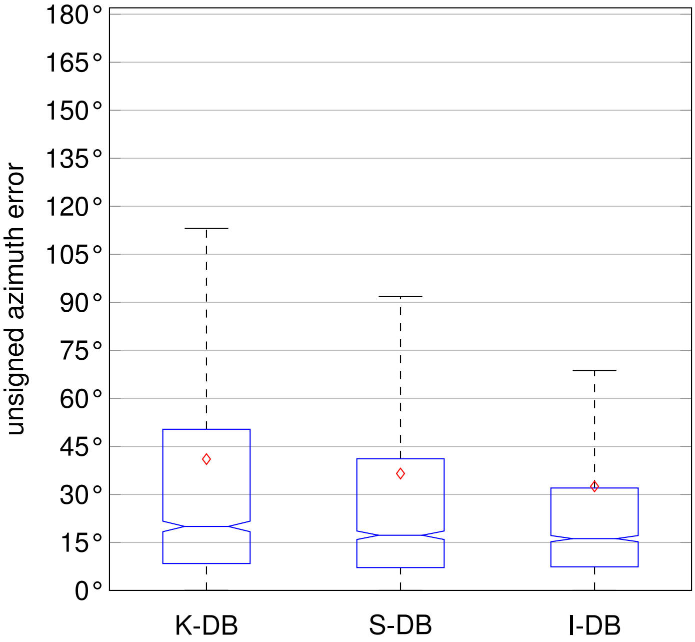

Figure 10.6.: SL (L1) evaluation of different sets of HRTFs without head tracking (NHT)

SL(L1): HRTF Dataset

An interesting question is whether the K-DB, the S-DB and the I-DB lead to different sound source localization results of the listeners. Due to the huge influence of the head tracking option, I split the presentation of the dataset comparison into two parts. The first part presents the results of the NHT option and the second part deals with the HQHT option of the playback system.

Figure 10.6 illustrates the result of the listening experiment with disabled head tracking (NHT option). The I-DB performs best with a mean azimuth error of 32.49° and a percentage of front-back and back-front confusions of 17.9 %. The S-DB is ranked second with a mean azimuth error of 36.52° and a confusion rate of 19.9 %. The K-DB is ranked last with a confusion rate of 24.5 % and a mean azimuth error of 41.02°. The differences of the I-DB and the K-DB are statistically significant ($p < 0.01$) as well as the differences between K-DB and S-DB ($p < 0.01$). The customized (S-DB) and the individual HRTF database (I-DB) show advantages regarding the sound localization ability of the probands in our SL test setup. Obviously the better representation of the HRTF's interaural time differences, level differences and spectral properties have significant influences of the sound localization performance in the NHT scenario. However, it is worth mentioning that the SL performance improvement of the S-DB and I-DB are achieved by extra effort for a previous selection or measurement process.

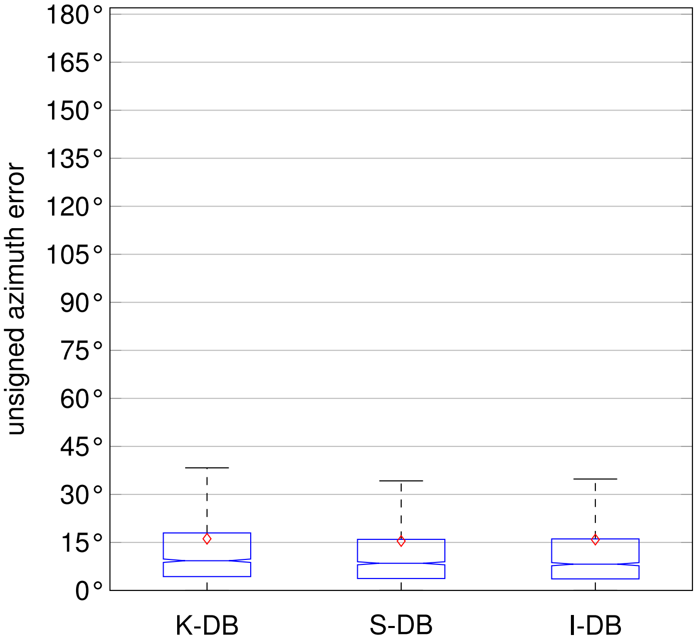

Figure 10.7.: SL (L1) evaluation of of different sets of HRTFs with enabled head tracking (HQHT)

The results with enabled head tracking (HQHT) are shown in Figure 10.7. The overall sound localization performance of the probands is dramatically better, considered that only little head movements of the listeners are possible due to the restricted length of the stimuli (2 s). The restricted stimuli lengths, however, are sufficient to decrease the number of front-back and back-front confusions to a big extent. The improved accuracy in terms of mean angular error of the sound source localization with HQHT is therefore mainly caused by a lower number of confusions, independent of the utilized HRTF dataset. The mean angular error for the I-DB is 15.89° and the percentage of confusions is 4.6 %. The S-DB reaches a mean angular error of 15.44° and a confusion rate of 4.2 %. The non-individual K-DB achieves a mean angular error of 16.14° with a confusion rate of 4.6 %. Interestingly, the sound localization differences between the non-individual K-DB, the customized S-DB and the individual I-DB vanish if head tracking is enabled. Thus, the minor differences in the SL results are not statistically significant.

Listening Test L2

In listening test L2 a second group of probands is confronted with the sound localization experiment. Besides the K-DB that is already used in L1, the conferees in L2 listen to stimuli that are virtually synthesized by use of the regression generated HRIR datasets (R-DBs). Since the L1 evaluation session unveiled that there are no significant differences

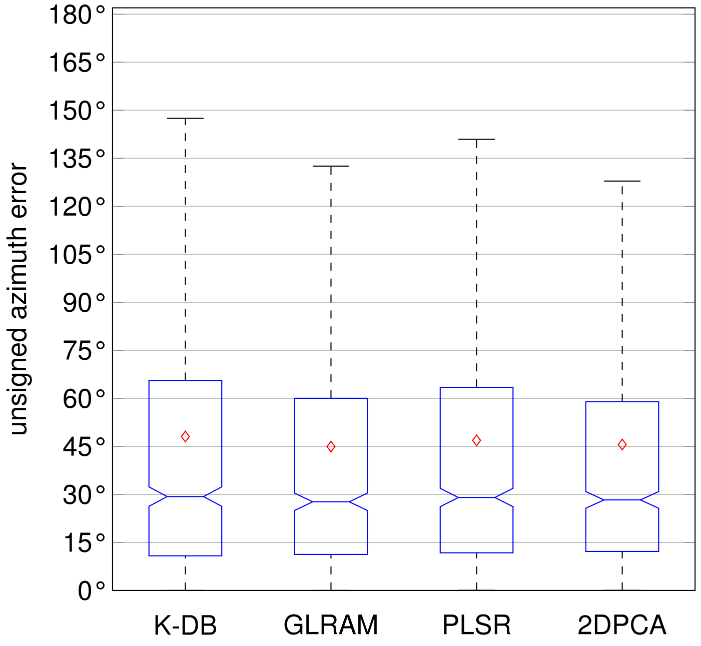

Figure 10.8.: SL (L2) evaluation of different sets of HRTFs without head tracking

between the noise burst and speech stimuli, we only use the speech stimulus in the L2 campaign, which reduces the time needed for the listening tests. The speech stimulus is also more representative for the teleconferencing scenario.

Figure 10.8 shows the evaluation results for the different regression generated HRTF datasets in NHT mode and 10.9 illustrates the results of the HQHT mode. As already observed in L1, there also is a huge difference in L2 between the HQHT-mode results and the NHT-mode performance, e.g., the mean angular error difference between the L2-session K-DB in NHT and the K-DB in HQHT is 27.43°. In the L1-session, the difference between HQHT and NHT with the K-DB is 24.88° which can be considered as a similar difference concerning the tracking options in both sessions.

For both tracking modes in L2, there are no statistically significant differences between the K-DB and the three regression computed datasets, namely, the GLRAM, the PLSR and the 2DPCA dataset. Taking the performance of the K-DB in L1 into account, the regression generated datasets can be considered slightly worse compared to the I-DB and the S-DB for the NHT mode and as good as the other datasets in the HQHT mode.

SL: Concluding Remarks

The sound localization (SL) campaign gives a first impression about the possible immersive playback options for a remote listener of the developed teleconferencing system. The

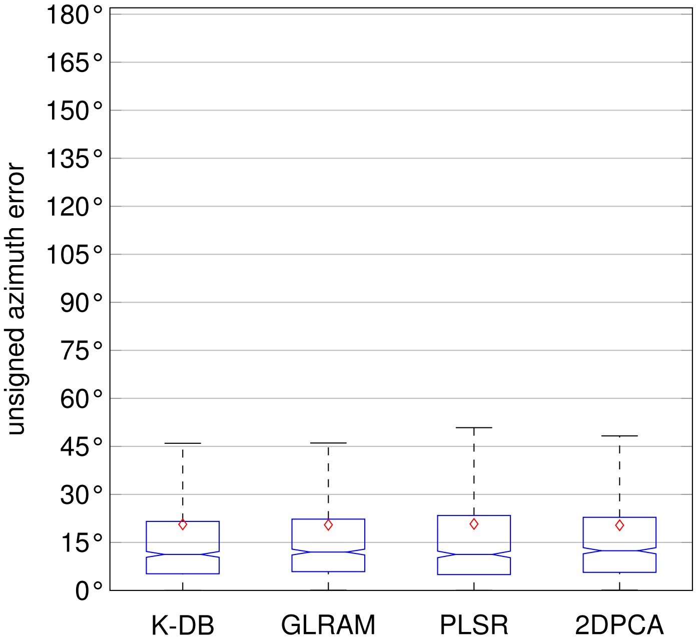

Figure 10.9.: SL (L2) evaluation of of different sets of HRTFs with enabled head tracking

enabled high quality head tracking (HQHT) has a big influence on the proband's localization performance and reduces the amount of front-back and back-front confusions significantly, compared to the disabled head tracking mode (NHT). The different datasets achieve similar performances in the HQHT mode which is expressed by the absence of any statistically significant difference between the dataset's evaluation results. In the NHT mode, statistically significant differences between the K-DB and the S-DB and between the K-DB and the I-DB can be observed, which indicates that the individual measured datasets and the selection process can improve the localization performance of the probands. The customization by regression does not indicate an improvement of localization accuracy compared to the usage of the K-DB. My research team and I provide a detailed description of the SL listening campaign and further results in [120].

The sound localization task is a frequently used measurement to judge the immersive sound presentation performance of HRTF datasets. However, the sound localization performance of a remote conferee does in my opinion not allow to draw conclusions about the user experienced performance of the developed teleconference system.

In the next chapter, an alternative concept, called Quality of Experience (QoE) is used to evaluate the sound acquisition of the teleconferencing system as well as the playback options of the developed system.

10.3. Quality of Experience Evaluation

After the SL experiments, the thesis proceeds with a more task-specific evaluation campaign of the developed teleconferencing system, the QoE evaluation.

Experimental Settings: QoE

The developed immersive teleconferencing system prototype consists of a microphone array that is placed on a conference table. The recorded speech contributions of the different conference participants are assigned to individual audio channels by the ALSR algorithm, described in Chapter 5.1. The assigned audio signals are finally virtually synthesized and played back to the experimentees via headphones.

Playback System Variations

Motivated by the findings of the SL evaluation a third head tracking option was developed and tested in the QoE listening tests. Since the HQHT mode significantly improved the SL results, I consider the use of head tracking desirable for the teleconferencing system. However, an expensive HQHT system is not available for many remote users of the teleconferencing system. Therefore, a low budget, low-quality head tracking (LQHT) is developed at the institute for data processing that seeks to enable dynamic sound synthesis by the use of a standard webcam [25] which is usually accessible for remote conferees. The utilized algorithm detects features (mouth, nose and eyes) in the listener's face and estimates the head pose of the remote conference participant. The drawbacks of the LQHT compared to the HQHT are illumination dependencies of the pose estimation, a restricted range in which the feature points of the face can be detected due to the direct line of sight requirement between the feature points and the webcam. The system recognizes azimuth angles between -35° and 35° and elevation angles between -10° and 40° at a frame rate of 23 fps [25].

Besides the LQHT, the SL playback system variations are used, namely, HQHT, NHT, K-DB, S-DB, I-DB and R-DBs.

Experimental Design

Similar to the SL evaluation, the QoE listening experiments consist of two parts, Q1 and Q2 which are conducted in separate sessions by two different groups of probands.

In Q1, the QoE evaluation listening experiments consist of channel assigned recordings of real world conference situations that are played back to the user with the afore mentioned options, namely, HQHT, LQHT and NHT in combination with K-DB, S-DB and I-DB.

In order to judge the assignment algorithm, there are two types of conference situations that are assigned and evaluated individually. The first situation denoted as CI consists of

Assignment	ALSR									HS								
Head tracking	HQHT			LQHT			NHT			HQHT			LQHT			NHT		
HRTF-DB	K	S	I	K	S	I	K	S	I	K	S	I	K	S	I	K	S	I

Table 10.3.: Q1, CI: Schematic overview of the different stimulus treatments for conference situation CI.

a common teleconference situation whereas the second situation, CII, describes a discussion with simultaneously active participants resulting in chaotic discussions.

An ideally assigned conference situation served as a reference for the channel assignment system, which can also be considered to be a realistic special case of our teleconference scenario. This special case can be achieved if all conference participants communicate via headsets, resulting in perfectly separated and assigned conference contributions. Therefore, this case will be referred to as headset scenario (HS). We avoid to present mono playback to serve as lower limit due to possible desensitization effects of the test subjects [79].

In Q2 another group of listeners has to judge the QoE of the teleconferencing system using the regression generated (R-DB) HRTF datasets for conferencing situation CI.

Stimulus Treatments

As illustrated in Table 10.3, there are 18 possible stimulus treatments for situation CI, which are tested in a full factorial, within-subject design, which is the same stimulus treatment method as in the SL evaluation.

For conference situation CII a subset of possible playback combinations is evaluated in order to quantify the benefit of spatially separated playback even if the assignment includes artifacts. Table 10.4 gives an overview about the stimulus treatments for CII. In addition to the stimulus treatments of CI, the virtual conferee placement is tested for CII. The virtual conferee placement denotes the spatial placement of the conference participants, 3D virtual conferee placement means that the four conference participants are virtually synthesized to four different positions, whereas Mono (M) conference conferee placement results in the virtual placement of all conference participants to one position in order to estimate the benefit of differently HRTF-synthesized playback positions compared to a mono-like conference playback.

Table 10.5 shows the stimulus treatments for Q2 for situation CI, where. The K-DB serves as a reference to indirectly compare the results, achieved with the regression HRTFs with the results of Q1 for situation CI.

Head tracking	HQHT			
HRTF-DB	K-DB			
Assignment.	ALSR		HS	
Virtual conferee placement	3D	M	3D	M

Table 10.4.: Q1, CII: Schematic overview of the different stimulus treatments for conference situation CII.

Assignment	ALSR				HS			
Head tracking	HQHT		NHT		HQHT		NHT	
HRTF-DB	K	R	K	R	K	R	K	R

Table 10.5.: Q2, CI: Schematic overview of the different stimulus treatments for conference situation CI. The HRTF-DB denoted by "R" consists of the regression generated 2DPCA, GLRAM and PLSR dataset.

Stimuli

To evaluate our teleconferencing system, speech sources serve as stimuli, which is also suggested in [124]. Furthermore, the tests are conducted by use of a meeting corpus that is designed for the purpose of evaluating teleconferencing situations [158].

The conference sound acquisition system is installed in accordance with the channel assignment evaluation experiment in echoic conditions as described in Section 5.2, where the conference participants of the conference corpus [158] are played back via four different loudspeakers in order to have reproducible recording sessions. The microphone array recordings are then processed by the ALSR channel assignment algorithm.

Each conference listening stimulus is further processed with the different stimulus treatments and has a length of 58 s which gives the test listeners enough time to put themselves into the conferencing participant position. The conference consists of four actively participating conferees. The ALSR algorithm is trained with a 10 s-introduction round which is consequently not part of the actual test material.

Stimuli Presentation Sequence and Scale

In order to avoid listening test results that are dependent on the stimuli presentation order, a balanced latin square design is applied to determine the stimuli presentation sequence.

The listeners can express their opinion about the presented stimulus with the Bodden Jekosch scale in accordance with [114] and [124]. The scale is extended at the scale's edges to avoid contraction bias as mentioned in Section 9.3.

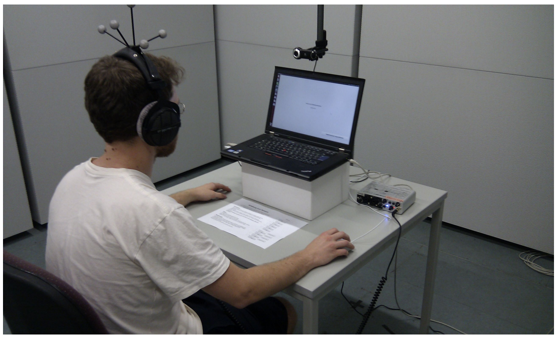

Figure 10.10.: A listener conducts the QoE evaluation in the LDV audiolab

QoE: Experimental Procedure

The QoE listening experiment is conducted at the LDV audiolab [145]. The HQHT and the LQHT determine the head movements simultaneously such that the proband can not differentiate between these tracking options besides the listening experience. We implement an interface to collect the listener's judgements. Figure 10.10 illustrates the test setup with the LQHT using the webcam and the HQHT using the markers that are fixed at the *Beyerdynamic DT 990 Pro* headphones. The laptop serves as an evaluation interface allowing the listener to conveniently judge the respective QoE. The laptop furthermore processes the sound samples and the head rotation according to the corresponding stimulus treatment. The stimuli are played back by a *Roland UA 25 EX* audio interface. Besides practical use, the experimental settings represent a tradeoff between reproducible listening test requirements and the teleconferencing scenario of a remote listener.

The determination of each listener's I-DB was done in a separate measurement meeting several weeks before the actual QoE evaluation. The individual selection of the S-DB was also made several days in advance of the actual QoE sessions.

At the beginning of the QoE tests, each proband receives a written briefing about the following training and test procedure. The training makes the proband familiar with the QoE test by presenting the key functionalities of the teleconferencing QoE test platform as follows. First, we seek to sensitize the probands to virtually synthesized playback of conferees by short conferencing examples. Second, the probands should get used to dynamic sound synthesis, therefore, a training example of a conference is presented by using HRTF-based sound synthesis in the HQHT mode. Third, the listener should listen to possible artifacts caused by the ALSR algorithm. Fourth, the graphical user interface and its handling to judge the QoE results of the listening tests is described. Please refer to [176] to access the written briefing for the QoE evaluation and for determining the S-DB.

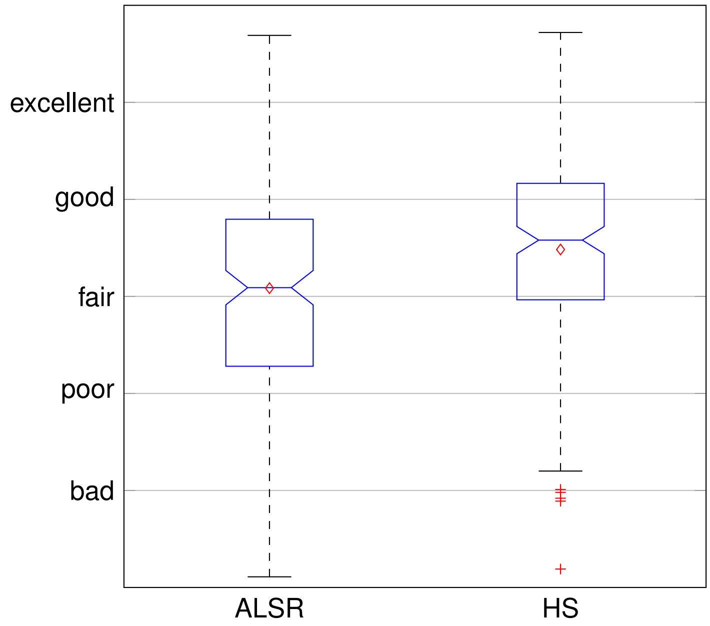

Figure 10.11.: QoE (Q1, CI) evaluation of ALSR assigned conference contributions compared to the ideally headset (HS) assigned conference

Experimental Results: QoE

This section presents the results of the QoE listening experiments that were conducted with 20 subjects. Similar to the SL evaluation, the results are presented with Boxplots and the statistical analysis is done by an analysis of variance (ANOVA). Again, we consider results with p-values lower than $p = 0.05$ as statistically significant.

QoE Q1: Conference Situation CI

The first part of the listening experiment presents the probands conference situation CI with different stimulus treatments summarized in Table 10.3.

Q1, CI: Assignment

Figure 10.11 illustrates the results of the conference recordings that are assigned with the ALSR algorithm compared to the ideal headset speaker assignment (HS). Of course, the ideally assigned HS stimuli yield better QoE judgements than the assignment of the microphone array recordings that are processed with the ALSR algorithm. The ANOVA provides a statistically significant QoE difference between the ALSR and HS assignment, expressed

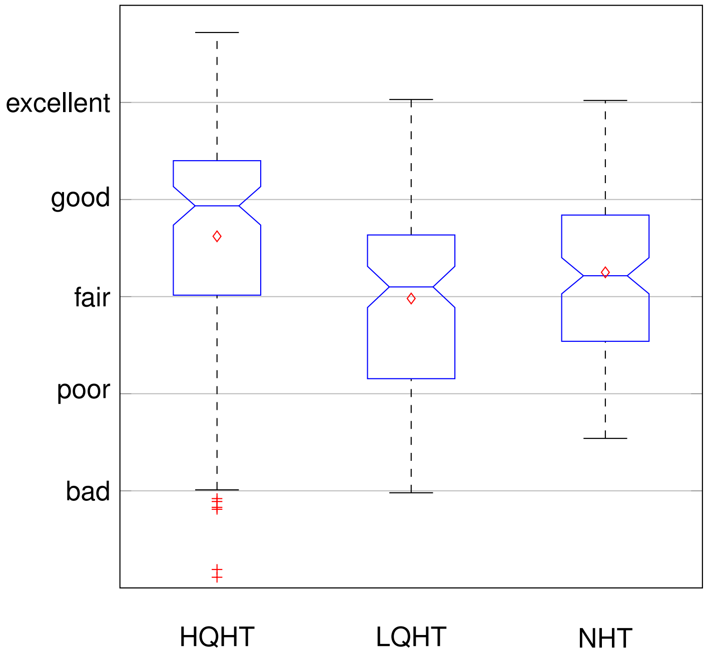

Figure 10.12.: QoE (Q1, CI) evaluation of the different head tracking options HQHT, LQHT and NHT.

by $p < 0.01$. Remarkably, the QoE-difference between the both entirely different assignment conditions is little, meaning that the developed microphone array in combination with the ALSR channel assignment works audibly well under realistic conference conditions in an echoic environment.

Q1, CI: Head Tracking

Another important question for the proposed teleconferencing system is whether it is valuable for the perceived QoE to equip the remote conferee with head tracking in order to enable dynamic sound synthesis. As illustrated in Figure 10.12, the conference presentation with the HQHT is superior to the NHT stimulus treatments. However, the LQHT is perceived worse than the NHT option. Due to $p < 0.01$ of the ANOVA, the differences are likely to be statistically significant. The thereupon conducted least significant difference (LSD) method [51] for pairwise comparison of the results eventually confirms the statistical significance of the perceived QoE differences with respect to the tracking options.

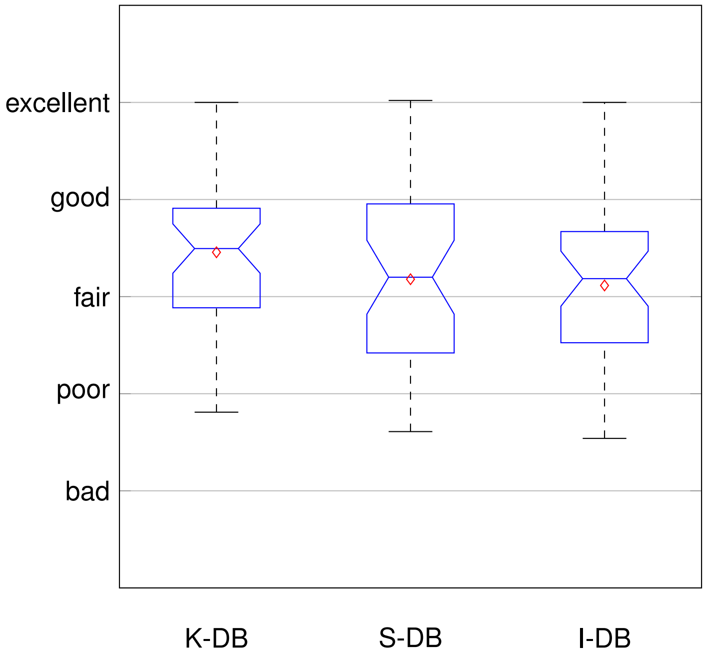

Figure 10.13.: QoE (Q1, C1) evaluation of the different HRTF datasets and disabled head tracking (NHT)

Q1, CI: HRTF Database

Finally, the user's QoE impression of the different available HRTF datasets, namely the K-DB, the S-DB and the I-DB, is analyzed in dependence of the head tracking mode for conference situation CI.

Figure 10.13 pictures the QoE judgements if the head tracking is disabled which results in a static HRTF-based synthesis of the conferee's contributions. According to the ANOVA, there is no significant difference between the different datasets that are used for the virtual placement of the conference participants. Surprisingly, the listeners do not clearly prefer their own I-DB. Moreover, there is a slight tendency toward the K-DB which is the only non-individual HRTF dataset in the tests, whereas the I-DB and the S-DB seek to address the different geometric features of the listeners.

In Figure 10.14 the listening test results for the dynamic conference sound synthesis using LQHT are illustrated. According to the ANOVA analysis, there is no significant difference between the stimuli generated by the different HRTF datasets. As already discussed, the perceived QoE is higher for the NHT system than for the LQHT system, which can also be observed by comparing the QoE judgements in Figures 10.13 and 10.14. Interestingly, the ranking of the different HRTF databases changes, though, not statistically significant. The QoE of the S-DB stimuli nearly stays at the same level for NHT and LQHT, whereas

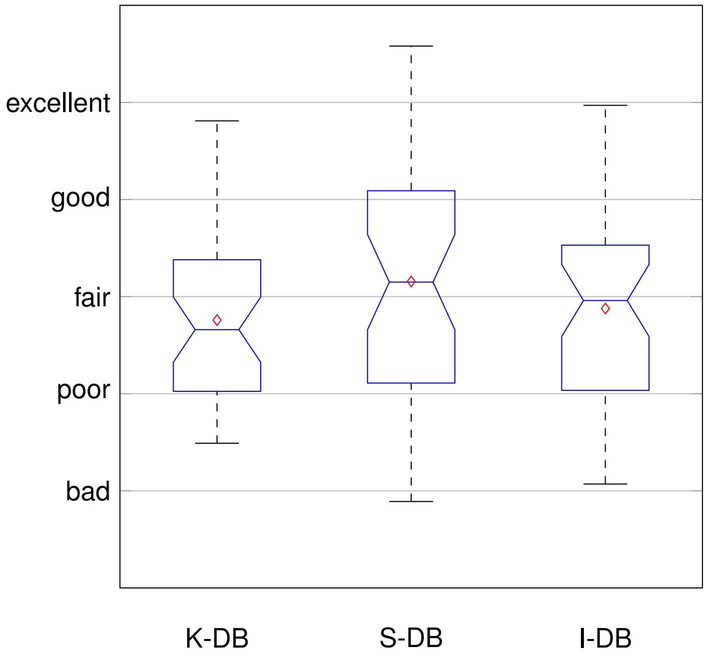

Figure 10.14.: QoE (Q1, CI) evaluation of the different HRTF datasets and enabled head tracking (LQHT)

the QoE of the I-DB marginally decreases and the K-DB suffers huge loss in the user's perceived QoE. Consequently, the S-DB performs best in the LQHT setup.

Besides the NHT and the LQHT, the HQHT option has been subject to QoE testing. The QoE results with HQHT are illustrated in Figure 10.15. According to the ANOVA and LSD analysis, there exists a statistically significant difference between the I-DB and the S-DB. The differences between the I-DB and K-DB as well as between the K-DB and S-DB are not statistically significant. In the HQHT setup, the teleconference situation with the individually measured I-DB that is obtained by a cumbersome and time-consuming procedure (Section 10.1) is ranked first by the probands followed by the K-DB and the S-DB.

Contrary to my expectation that the I-DB would be preferred by most probands, the QoE evaluation reveals three different rankings of the available HRTF datasets in dependence of the head tracking mode of the teleconferencing system. In the NHT system, the K-DB performes best according to the listener's judgement. The S-DB was the probands favorite in the LQHT scenario and the I-DB was the users first choice in the HQHT mode. The listening experiment revealed only minor differences among the eligible HRTF datasets. Interestingly, the head tracking option has an influence not only on the overall quality opinion but also on the ranking of the HRTF datasets. Furthermore, it is noticeable that with a more realistic head tracking option, the QoE impression using the I-DB improves com-

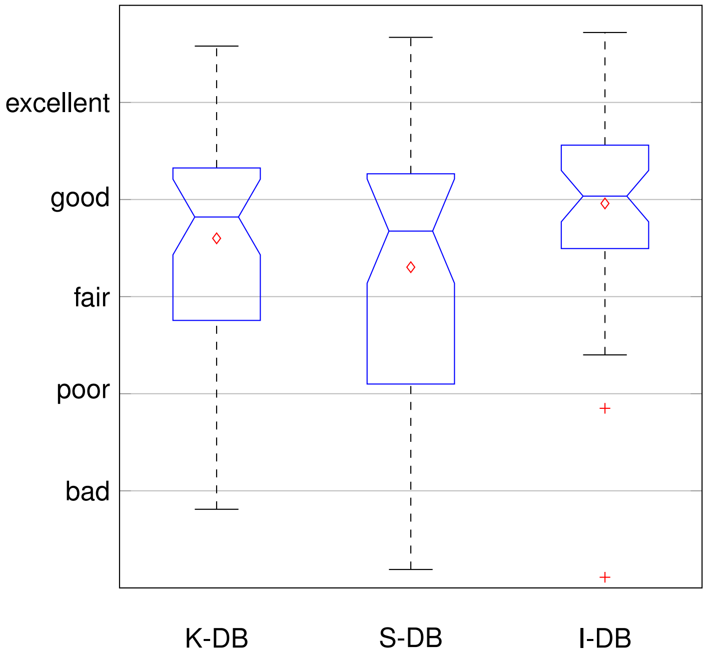

Figure 10.15.: QoE (Q1, CI) evaluation of the different HRTF datasets and enabled head tracking (HQHT)

pared to the two other datasets. It might be that the probands indeed prefer their own I-DB, provided that the head rotation can be precisely and quickly determined.

Our QoE results indicate that it might be even counterproductive to utilize I-DBs without head tracking or in combination with low quality head tracking for the virtual placement of conferees. Possibly the discrepancy between highly accurate head tracking and non-individual HRTFs produce slight confusions where the test subjects tend to rate the perceived QoE lower than for the HQHT and I-DB combination of the test set.

The same applies vice versa to the NHT and LQHT modes combined with the I-DB, where the realistic individually measured HRTF dataset does not fit to the non-realistic feeling of the LQHT or the NHT modes.

QoE, Q1: Conference Situation CII

The second part of the teleconferencing system QoE evaluation deals with conference situation CII. The conference situation is designed to state the worst case of a conference for the ALSR algorithm with simultaneously speaking conferees which causes channel assignment errors and separation artifacts. Table 10.4 gives an overview about the stimulus treatments for conference situation CII.

Figure 10.16 illustrates the listener's QoE judgment for conference situation CII. The

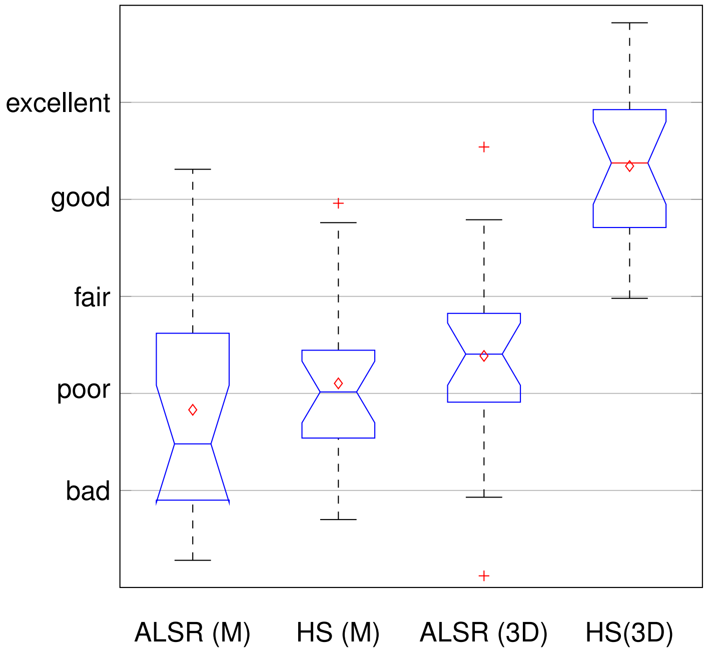

Figure 10.16.: QoE (Q1, CII) evaluation of the worst case scenario with simultaneously talking conferees. HS(3D) denotes a virtual separated placement of perfectly headset assigned conference participants, HS(M) describes a mono-like playback of perfectly assigned conferees. The ALSR option denotes an assignment of the microphone array recordings with the ALSR algorithm.

ANOVA and LSD analysis of the listening test results confirm a statistically significant difference between the HS/3D combination and the other three playback combinations. It is obvious that the ideally assigned and HRTF-synthesized separated speech contributions of the HS scenario outperform the other three options.

The distinctions of the other three stimulus treatments are not statistically significant according to the LSD analysis and have to be considered as tendencies. Interestingly, the probands still prefer HRTF-synthesized separated conferees that are assigned with the ALSR algorithm to the not separated conference placement stimuli. Due to the challenging conference situation, the ALSR algorithm does not perfectly assign the simultaneously active conferees resulting in audible artifacts. However, the virtual conferee placement to different separated positions is still favored to the mono-like non-separated placement.

QoE, Q2: Conference Situation CI

After session Q1, another group of test subjects judges the QoE of the regression-generated HRIRs for virtual playback within a teleconference. Based on the findings in Q1, we resign the LQHT option and conference situation CII in the Q2 scenario in order to

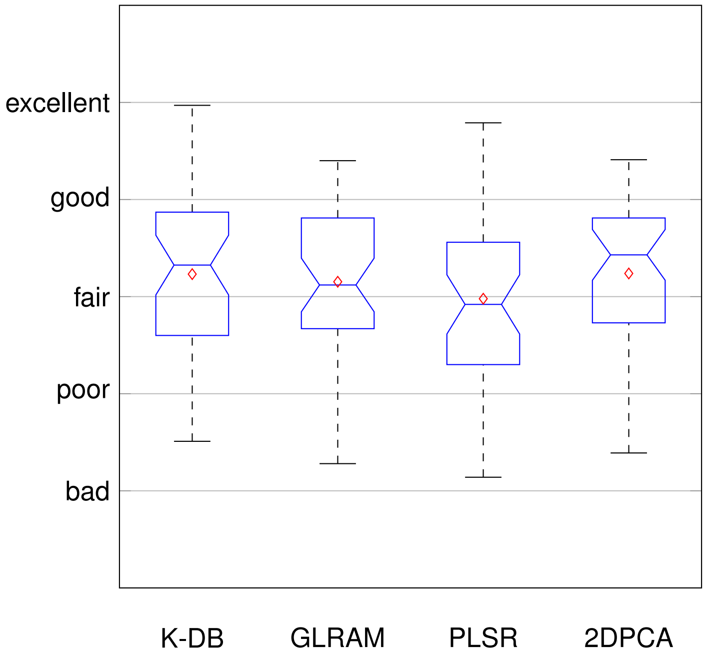

Figure 10.17.: QoE (Q2, CI) evaluation of the conference situation CI with disabled head tracking (NHT)

concentrate on the differences within the regression datasets (2DPCA, PLSR, GLRAM). Besides the regression datasets, the acoustically measured K-DB is also used in the experiment which allows an indirect comparison of the I-DB and the S-DB of Q1 with the results that the regression datasets yield in Q2.

Figure 10.17 shows the QoE evaluation results for the different regression-generated datasets. As already mentioned, the listener's regression datasets are computed based on a few anthropometric measurements. Using the NHT option, there is no statistically significant difference between the regression datasets, denoting that the minor spectral distortion (SD) differences listed in Section 7.3 do not result in QoE distinctions.

The same applies to the Q2 evaluation with enabled HQHT. There are no statistically significant differences between the regression-generated individual datasets, as seen in Figure 10.18. Similar to session Q1, the overall QoE-ratings with enabled HQHT are higher than with disabled head tracking and the overall ratings in Q2 are similar to the Q1, C1 session looking at the K-DB results. In both tracking options, the non-individual K-DB does not show worse QoE results than the computed individual regression generated datasets.

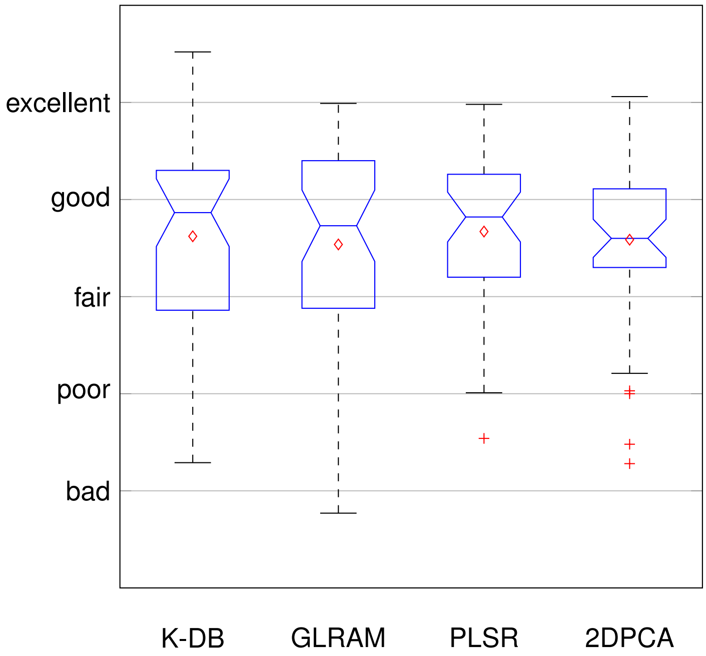

Figure 10.18.: QoE (Q2, CI) evaluation of the conference situation CI with enabled high quality head tracking (HQHT)

QoE: Concluding Remarks

The QoE listening campaign offers valuable clues to the benefits of the different applied techniques to achieve an immersive teleconference impression for a remote conferee. Especially the different possible sound synthesis options that have evolved as part of this work have been tested in detail for the teleconference scenario.

It has been shown that the users of our teleconferencing system appreciate dynamic sound synthesis, provided that the applied head tracking is of high quality. Even if not consciously perceived by the probands, a webcam based low quality head tracking solution is judged worse than the static playback of the conference contribution without the use of head tracking.

Moreover, the listening tests unveil that an individually measured HRTF database does not add a leap in QoE in the context of teleconferencing. There are tendencies that individually measured HRTF databases are preferred in combination with HQHT, but without head tracking, the non-individual K-DB performs even slightly better as the costly measured I-DB. The regression-computed datasets based on individual anthropometric measurement do also show no significant QoE-improvement compared to the K-DB in the teleconference scenario.

Furthermore, it has been shown that the developed ALSR algorithm for channel assignment works audibly well for the processing of the microphone array recordings in realistic and echoic conference conditions. In addition, immersive playback of conferees is preferred even if the sound acquisition and speaker assignment system produces artifacts due to simultaneously active conference participants.

Unfortunately, a HQHT system is nowadays not accessible to the most remote teleconference participants using smartphones or a standard computer and the listening experiment showed that the developed LQHT is not an adequate substitute for the HQHT in terms of QoE. Therefore, I choose to use the NHT option for further cognitive load (CL) investigations. Also, the QoE judgements do not clearly favor any certain HRTF dataset. Consequently, using the K-DB seems appropriate to me due to the good QoE results and the fact that the K-DB is straightforwardly accessible to any remote teleconference participant in contrast to the other QoE-tested HRTF datasets. My research team and I provide more detailed information about the QoE campaign in [176].

10.4. Cognitive Load Evaluation

In this section, the cognitive load (CL) expended by the teleconference participants is evaluated while using the developed teleconferencing system.

Experimental Settings: CL

The CL-evaluations are conducted in the same setup as the QoE tests. The microphone array records conference speech contributions presented by loudspeakers in an echoic environment and the ALSR algorithm assigns the speech contributions to individual channels that are presented to the proband via headphones. Due to the findings in the previous QoE-evaluation campaign, the K-DB and the NHT option was used in the CL-tests.

To benchmark the benefit of binaural playback (CL-B) in our teleconference system, we extend the stimulus treatments by also presenting the conferences in mono (CL-M) to the test listeners. The stimuli presentation sequence (latin square) and the utilized scale (Bodden Jekosch scale) are in accordance with the QoE tests.

The used speech contributions played back with the different loudspeakers around the system's microphone array are the conferences which were carefully designed and recorded by the Telekom Innovation Laboratories [158]. In contrast to the QoE tests, the CL investigation demands different conferences each with the same level of difficulty regarding the conference contents, since the probands of the CL evaluation must not have any learning effect between the different presented stimulus treatments. The utilized conference corpus [158] is designed to fulfill this requirement.

CL: Experimental Procedure

After an introduction, the probands are confronted with six different conferences out of the conference corpus. Three conferences are held by three conferees and three conferences consist of four contributing participants. Each proband has to listen to three mono recordings of the conference and three HRTF-synthesized conference recordings which are recorded with the afore mentioned microphone array and processed by the ALSR algorithm. The presentation sequence, the binaural and mono layout of the conferences is different for each proband according to a latin square design. The conferences last between five and nine minutes. Directly after listening to each of the six conferences, the probands have to complete questions regarding the conference content, e.g., answer a question who contributed a certain information during the conference. Furthermore, the listeners have to judge the experienced level of assurance for each answer. Moreover, the probands have to estimate the amount of listening effort needed to follow the conference and the degree of difficulty in identifying the talking person within a conference.

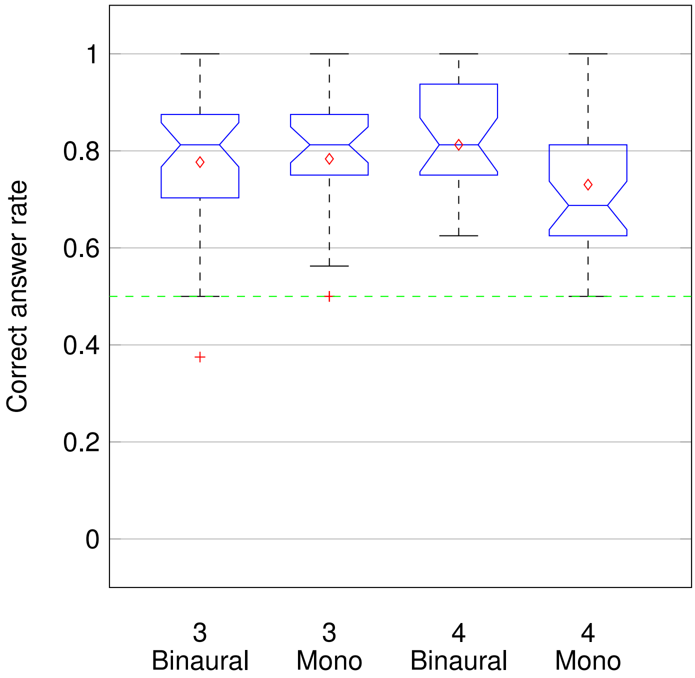

Figure 10.19.: CL evaluation of the memorized conference contents with three and four conference participants [103]

Experimental Results: CL

The CL listening experiments are conducted with 21 probands. In accordance with the QoE evaluation in Section 10.3 the ANOVA is applied to determine statistically significant differences between the stimulus treatments.

Memorizing Conference Contents

The most objective criterion to evaluate different playback options in the conference scenario is to ask conferees questions about the facts discussed in a conference meeting. Therefore, after each conference session the probands answer 16 questions which seek to determine the amount of memorized conference content.

Figure 10.19 illustrates the amount of correct answers for the different stimulus treatments. For conferences consisting of three conferees, there is no significant difference between the virtually HRTF-synthesized binaural playback (CL-B) and the mono option (CL-M). Slightly more than 80 % of questions were answered correctly. For conferences with four participants, a significant difference with a $p = 0.011$ between the CL-B and CL-M playback can be observed. Using the binaural playback the percentage of correctly answered questions is 8.2 % higher than for the mono conference scenario.

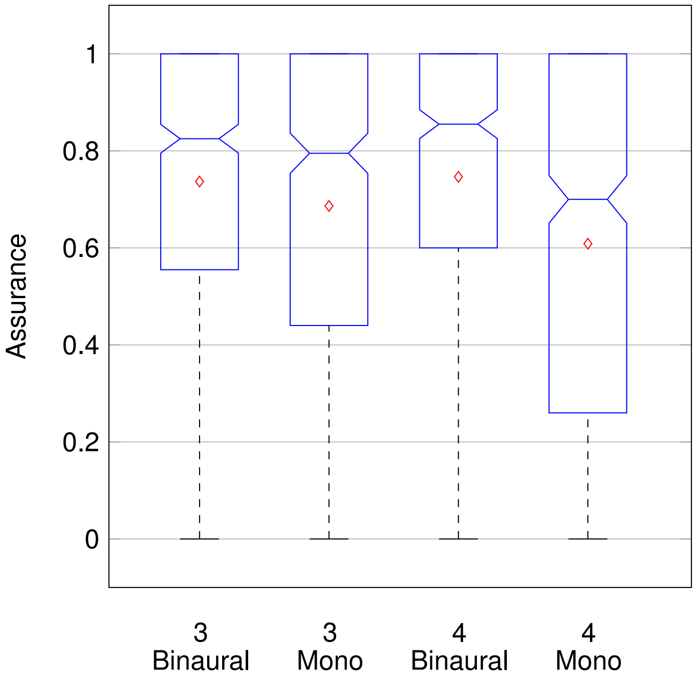

Figure 10.20.: CL evaluation of the perceived assurance regarding the answered teleconference questions [103]

The evaluation results show that the user of the developed immersive teleconference system objectively benefits from a higher rate of memorized conference content for four conferees. Therefore, the aim of effective communication with a teleconferencing system is supported by the possibility to offer 3D playback of conference contributions which requires a preceding channel assignment.

Assurance

Besides the objective determination of the memorized conference contents, the probands are asked about the degree of assurance of their answers in the memorizing task. Figure 10.20 shows that the CL-B scenario is superior to the CL-M scenario. Both p-values indicate a statistical significance between Cl-M and Cl-B, however, it is worth mentioning that the probands are instructed to truly mark a certainty level of zero if they do not have any idea about the true answer which possibly influences the assumption of normally distributed answers. Again the difference between CL-M and CL-B is higher for the conference with four participants than the conference with three participants.

This second CL experiment unveils that not only the percentage of correct answers is

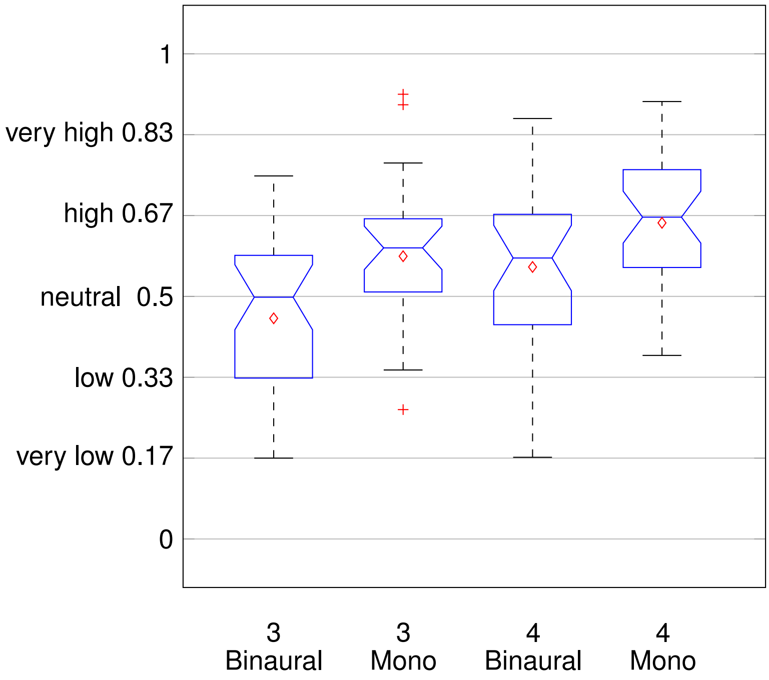

Figure 10.21.: CL evaluation of the subjectively perceived effort to concentrate on the teleconference situation [103]

higher in the CL-B situation but also the proband's certainty about the given answers is higher in the CL-B conferences compared to the CL-M scenario.

Listening Effort

Another question besides the effectiveness of conferences, pictured by the memorized content and the corresponding degree of assurance, is the efficiency of a conference discussion, e.g. given by the amount of concentration that is needed to follow a conference conversation. Therefore, the probands are asked to estimate the amount of listening effort that was needed to follow the conference. The evaluation results are illustrated in Figure 10.21, where a significant difference between the mono and binaural playback is observed with statistically significant p-values of $p < 0.01$ for the three conferee scenario and $p = 0.029$ for the conference held by four participants.

In sum, the probands indicate that a higher amount of effort is needed to follow a conference consisting of four conferees and the listening effort is always lower in the CL-B conference than in the CL-M conference. It is worth mentioning that with the CL-B technique, the additional conference participant in the four conferee scenario can be compensated compared to the three conferee scenario.

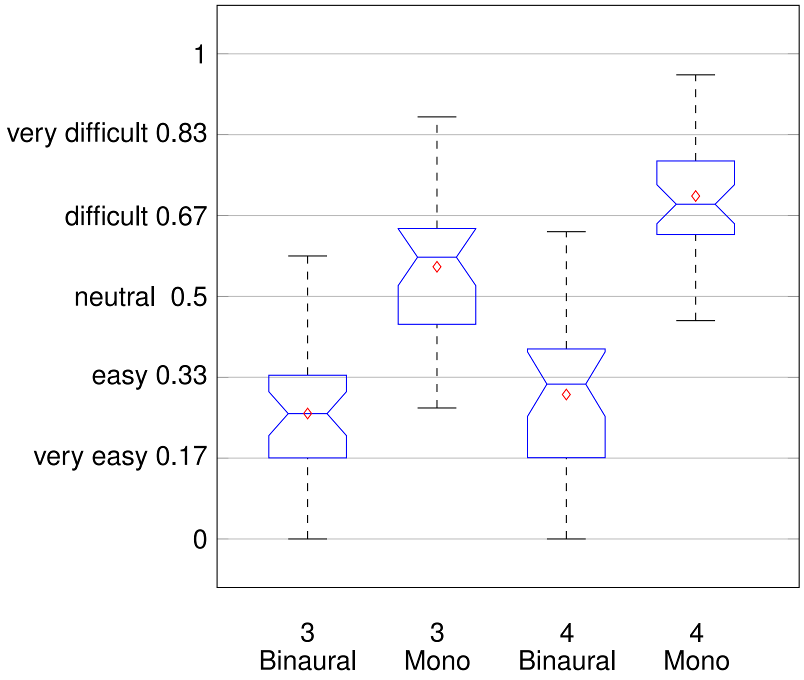

Figure 10.22.: CL evaluation of the subjectively perceived effort to identify the active conferee [103]

Speaker Identification Effort

Finally, the probands subjectively judge the effort needed to identify the active conferee. Figure 10.22 shows that the speech signal and the additionally available direction information in the CL-B playbacks greatly improve the proband's judgement about their ability to differentiate between the conference participants. The improved ability to identify the active conference participant when using the CL-B option could also influence the memorization task in a positive way.

CL: Concluding remarks

Besides the QoE, the cognitive load was identified to serve as a quality criterion for the developed teleconferencing system. The CL-listening test unveils objective and subjective differences between an immersive (CL-B) and a conventionally (CL-M) conducted teleconference for a remote participant.

The evaluation campaign shows that the developed immersive teleconferencing system allows a better performance in memorization of conference contents, which states a huge success for the developed system since the memorization of information exchange during a conference is the most primary reason to conduct conferences. The advantage of the memorization task concerning the CL-B playback is significant for four conferees whereas there are small differences between CL-B and CL-M for three participants.

Besides the objective task of keeping conference discussion facts in mind, the assurance level of the probands in the immersive scenario is higher than in the CL-M scenario. Furthermore, test participants feel that the required amount of listening effort and the speaker identification effort to follow the immersive CL-B conferences is less than in the CL-M scenario. In [103], we provide detailed information about the cognitive load listening tests.

11. Conclusion

In this thesis, I presented the development of an audio conferencing system with respect to a number of requirements. In the beginning of this work three issues at the system level, concerning the sound acquisition, the immersive playback and the system evaluation, were raised which were answered by numerous experiments and listening tests.

At the sound acquisition site, various sound source localization and sound source separation algorithms were benchmarked for differently shaped microphone array prototypes that were designed with respect to the different algorithms. The comprehensive experiments in echoic and anechoic environments suggested that a sound source localization algorithm, called steered response power - phase transform (SRP-PHAT) algorithm in combination with the geometric source separation (GSS) algorithm, provides a promising basis for the sound acquisition system. Beside sound source localization and separation, an online speaker recognition algorithm based on Gaussian mixture models, was investigated to additionally include the conferee's voice features into the assignment considerations. The complementary strengths and drawbacks of the localization and separation algorithm on the one hand, and the speaker recognition algorithm on the other hand, are finally combined to a sound acquisition system that fulfills the stated requirements for the conference system's sound acquisition. Experiments showed that the assignment algorithm is able to reliably assign the active conferee to the individual channel, even in the case when two participants of the conference swap places.

At the immersive playback site, three different methods to enable a user-specific customizeable HRTF-based sound synthesis were investigated, namely the selection, the regression and the acoustic measurement method. The three methods offer a different degree of HRTF-individualization which is achieved by different levels of effort. For each method, several specific approaches were experimentally benchmarked.

To compare the different individualization methods for the immersive conference playback by listening tests, it was necessary to set up a whole HRTF database with anthropometric data which also includes the HRTFs of the test subjects. Among the different acoustic measurement approaches, a recent approach using a continuous recording of the excitation signal provides the best cost-benefit-ratio in the experiments and was therefore used to set up the LDV HRTF database.

Finally, an extensive evaluation of the developed conferencing system was conducted. Three evaluation concepts were identified to rate the different modules and playback options of the conferencing system: the sound localization, the quality of experience and the cognitive load concept. For each concept, an appropriate test environment was built and in sum 126 hours of listening tests were conducted.

The sound localization tests unveil that the different HRTF datasets passed the listening tests almost equally with slight advantages for the acoustically measured individual dataset if head tracking was disabled. The quality of experience evaluation, which is more appropriate to the conferencing scenario, showed that the channel assignment algorithm works audibly well and that no statistically significant differences exist between the perceived quality of experience for the individualized datasets compared to the non-individualized set of HRTFs. The cognitive load evaluation finally proved that the developed assignment system, in combination with the non-individualized immersive playback system, increases the effectiveness of a conference. This is expressed by a higher amount of memorized conference content compared to the traditional mono playback of the conference. Furthermore, efficiency is increased denoted by a subjectively judged lower effort to identify the active conferee and a higher assurance level about answered questions after the conference.

Bibliography

1. F. Abrard and Y. Deville. A Time-Frequency Blind Signal Separation Method Applicable to Underdetermined Mixtures of Dependent Sources. In *Signal Processing*, 85(7), pp. 1389–1403, 2005.

2. V. Algazi, C. Avendano, and D. Thompson. Dependence of Subject and Measurement Position in Binaural Signal Acquisition. In *Journal of the Audio Engineering Society*, 47(11), 1999.

3. V. Algazi, R. Duda, D. Thompson, and C. Avendano. The CIPIC HRTF Database. In *In Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 21–24. 2001.

4. AMI Consortium. The AMI Meeting Corpus. URL `http://corpus.amiproject.org`. Accessed: 23.07.2013.

5. S. Araki, M. Fujimoto, K. Ishizuka, H. Sawada, and S. Makino. A DOA Based Speaker Diarization System for Real Meetings. In *In Proceedings of the Workshop on Hands-Free Speech Communication and Microphone Arrays*, pp. 29–32. 2008.

6. S. Araki, A. Ozerov, V. Gowreesunker, H. Sawada, F. Theis, G. Nolte, D. Lutter, and N. Duong. The 2010 Signal Separation Evaluation Campaign: Audio Source Separation. In *Latent Variable Analysis and Signal Separation*, pp. 114–122. Springer, 2010.

7. B. Arons. A Review of the Cocktail Party Effect. In *Journal of the American Voice I/O Society*, 12(7), pp. 35–50, 1992.

8. A. Arthur, R. Lunsford, M. Wesson, and S. Oviatt. Prototyping Novel Collaborative Multimodal Systems: Simulation, Data Collection and Analysis Tools for the Next Decade. In *Proceedings of the International Conference on Multimodal Interfaces*, pp. 209–216. 2006.

9. F. Asano, S. Ikeda, M. Ogawa, H. Asoh, and N. Kitawaki. Combined Approach of Array Processing and Independent Component Analysis for Blind Separation of Acoustic Signals. In *IEEE Transactions on Speech and Audio Processing*, 11(3), pp. 204 – 215, 2003.

10. A. Baddeley. Working Memory and Language: An Overview. In *Journal of Communication Disorders*, 36(3), pp. 189–208, 2003.

11. J. Baldis. Effects of Spatial Audio on Memory, Comprehension, and Preference During Desktop Conferences. In *Proceedings of the Conference on Human Factors in Computing Systems*, p. 166–173. 2001.

12. S. Bech and N. Zacharov. *Perceptual Audio Evaluation - Theory, Method and Application*. John Wiley & Sons, 2007.

13. D. Begault, E. Wenzel, and M. Anderson. Direct Comparison of the Impact of Head Tracking, Reverberation and Individualized Head-Related Transfer Functions on the Spatial Perception of a Virtual Speech Source. In *Journal of the Audio Engineering Society*, 49(10), pp. 904–916, 2001.

14. H. Beigi. *Fundamentals of Speaker Recognition*. Springer, 2011.

15. J. Berg and F. Rumsey. Systematic Evaluation of Perceived Spatial Quality. In *Audio Engineering Society Conference: Multichannel Audio, The New Reality*. 2003.

16. J. Blauert. *Spatial Hearing: the Psychophysics of Human Sound Localization*. MIT Press, 1997.

17. K. Blum, G. Van Rooyen, and H. Engelbrecht. Spatial Audio to Assist Speaker Identification in Telephony. In *International Conference on Systems, Signals and Image Processing*. 2011.

18. J. Boley and M. Lester. Statistical Analysis of ABX Results Using Signal Detection Theory. In *the Audio Engineering Society Convention*. 2009.

19. R. Bolia, W. Nelson, M. Ericson, and B. Simpson. A Speech Corpus for Multitalker Communications Research. In *The Journal of the Acoustical Society of America*, 107(2), pp. 1065–1066, 2000.

20. J. Bradley. Complete Counterbalancing of Immediate Sequential Effects in a Latin Square Design. In *Journal of the American Statistical Association*, 53(282), pp. 525–528, 1958.

21. M. Brandstein and H. Silverman. A Practical Methodology for Speech Source Localization With Microphone Arrays. In *Computer Speech & Language*, 11(2), pp. 91–126, 1997.

22. M. Brandstein and D. Ward. *Microphone Arrays: Signal Processing Techniques and Applications*. Springer, 2001.

23. A. Bronkhorst and R. Plomp. Effect of Multiple Speechlike Maskers on Binaural Speech Recognition in Normal and Impaired Hearing. In *The Journal of the Acoustical Society of America*, 92(6), pp. 3132–3139, 1992.

24. D. Brungart and B. Simpson. *Improving Multitalker Speech Communication With Advanced Audio Displays*. Technical Report, Air Force Research Lab Wright-Patterson Air Force Base Ohio, 2005.

25. P. Burger, M. Rothbucher, and K. Diepold. *Self-Initializing Head Pose Estimation With a 2D Monocular USB Camera*. Technical Report, Institute for Data Processing, Technische Universität München, 2014.

26. C. Busso, P. Georgiou, and S. Narayanan. Real-Time Monitoring of Participants' Interaction in a Meeting Using Audio-Visual Sensors. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, volume 2, pp. 685–688. 2007.

27. B. Cain. *A Review of the Mental Workload Literature*. Technical Report, Defence Research and Development, Human System Integration Section Toronto, 2007.

28. P. Callet, S. Möller, and A. Perkis. Qualinet White Paper on Definitions of Quality of Experience. European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), 2012.

29. W. Campbell, J. Campbell, D. Reynolds, E. Singer, and P. Torres-Carrasquillo. Support Vector Machines for Speaker and Language Recognition. In *Computer Speech & Language*, 20(2), pp. 210–229, 2006.

30. J. Cardoso. Multidimensional Independent Component Analysis. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, volume 4, pp. 1941–1944. 1998.

31. J. Cardoso. High-Order Contrasts for Independent Component Analysis. In *Neural Computation*, 11(1), pp. 157–192, 1999.

32. M. Casey and A. Westner. Separation of Mixed Audio Sources by Independent Subspace Analysis. In *Proceedings of the International Computer Music Conference*, pp. 154–161, 2000.

33. O. Cetin and E. Schriberg. Speaker Overlaps and ASR Errors in Meetings: Effects Before, During and After the Overlap. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, volume 1, pp. 14–19. 2006.

34. F. Chen. Localization of 3-D Sound Presented through Headphone-Duration of Sound Presentation and Localization Accuracy. In *Journal of the Audio Engineering Society*, 51(12), pp. 1163–1171, 2003.

35. E. Choueiri. *Optimal Crosstalk Cancellation for Binaural Audio with Two Loudspeakers*. Technical Report, 3D Audio and Applied Acoustics Laboratory, Princeton University, 2008.

36. P. Comon. Independent Component Analysis, A New Concept? In *Signal Processing*, 36(3), pp. 287–314, 1994.

37. J. Delosme, M. Morf, and B. Friedlander. Source Location from Time Differences of Arrival: Identifiability and Estimation. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pp. 818 – 824. 1980.

38. A. Dempster, N. Laird, and D. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. In *Journal of the Royal Statistical Society, Series B*, 39(1), pp. 1–38, 1977.

39. J. DiBiase, H. Silverman, and M. Brandstein. Robust Localization in Reverberant Rooms. In *Microphone Arrays*, pp. 157–180. Springer, 2001.

40. J. Dmochowski, J. Benesty, and S. Affes. On Spatial Aliasing in Microphone Arrays. In *IEEE Transactions on Signal Processing*, 57(4), pp. 1383–1395, 2009.

41. R. Drullman and A. Bronkhorst. Multichannel Speech Intelligibility and Talker Recognition Using Monaural, Binaural, and Three-Dimensional Auditory Presentation. In *The Journal of the Acoustical Society of America*, 107(4), pp. 2224–2235, 2000.

42. R. Duda and P. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, 1973.

43. M. Durkovic. *Localization, Tracking, and Separation of Sound Sources for Cognitive Robots*. Ph.D. thesis, Technische Universität München, 2012.

44. M. Durkovic, T. Habigt, M. Rothbucher, and K. Diepold. Low Latency Localization of Multiple Sound Sources in Reverberant Environments. In *The Journal of the Acoustical Society of America*, 130(6), pp. EL392–EL398, 2011.

45. G. Enzner. 3D-Continuous-Azimuth Acquisition of Head-Related Impulse Responses Using Multi-Channel Adaptive Filtering. In *Proceedings of the Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 325 – 328. 2009.

46. G. Enzner. Analysis and Optimal Control of LMS-Type Adaptive Filtering for Continuous-Azimuth Acquisition of Head Related Impulse Responses. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, volume 2, pp. 393–396. 2008.

47. N. Epain and J. Daniel. Pushing the Frequency Limits of Spherical Microphone Arrays. In *Proceedings of the Workshop on Hands-Free Speech Communication and Microphone Arrays*, pp. 9–12. 2008.

48. A. Farina. Simultaneous Measurement of Impulse Response and Distortion with a Swept-Sine Technique. In *the Audio Engineering Society Convention*, pp. 18–22. 2000.

49. H. Fastl. The Psychoacoustics of Sound-Quality Evaluation. In *Acta Acustica united with Acustica*, 83(5), pp. 754–764, 1997.

50. J. Feldmaier, M. Rothbucher, and K. Diepold. *Sound Localization and Separation for Teleconferencing Systems*. Technical Report, Institute for Data Processing, Technische Universität München, 2011.

51. A. Field. *Discovering Statistics using SPSS*. Sage Publications, 2009.

52. M. Frens, A. van Opstal, and R. van der Willigen. Spatial and Temporal Factors Determine Auditory-Visual Interactions in Human Saccadic Eye Movements. In *Perception & Psychophysics*, 57(6), pp. 802–816, 1995.

53. H. Fuchs. *Schallabsorber und Schalldämpfer*. Springer, 2009.

54. W. Gardner and K. Martin. *HRTF Measurements of a KEMAR Dummy-Head Microphone*. Technical Report, Massachusetts Institute of Technology, 1994.

55. J. Garofolo. *The 2009 (RT-09) Rich Transcription Meeting Recognition Evaluation Plan*. Technical Report, National Institute of Standards and Technology, 2009.

56. J. Geiger, F. Wallhoff, and G. Rigoll. GMM-UBM Based Open-Set Online Speaker Diarization. In *Proceedings of Interspeech*, pp. 2330–2333. 2010.

57. T. Grasser, M. Rothbucher, and K. Diepold. *Auswahlverfahren für HRTFs zur 3D Sound Synthese*. Technical Report, Institute for Data Processing, Technische Universität München, 2014.

58. T. Grasser, M. Rothbucher, and K. Diepold. *Speaker Localization and Separation in Teleconferences*. Technical Report, Institute for Data Processing, Technische Universität München, 2014.

59. G. Grindlay and M. Vasilescu. A Multilinear (Tensor) Framework for HRTF Analysis and Synthesis. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, volume 1, pp. 161–164. 2007.

60. F. Grondin and F. Michaud. WISS, a Speaker Identification System for Mobile Robots. In *Proceedings of the International Conference on Robotics and Automation*, pp. 1817–1822. 2012.

61. D. Hammershøi and H. Møller. Sound Transmission to and Within the Human Ear Canal. In *The Journal of the Acoustical Society of America*, 100(1), pp. 408–427, 1996.

62. J. Hao, I. Lee, T. Lee, and T. Sejnowski. Independent Vector Analysis for Source Separation Using a Mixture of Gaussians Prior. In *Neural Computation*, 22(6), pp. 1646–1673, 2010.

63. S. Hart. NASA-task load index (NASA-TLX); 20 Years Later. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 50, p. 904–908. 2006.

64. S. Hart and L. Staveland. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, 1(3), p. 139–183, 1988.

65. V. Hautamaki, T. Kinnunen, I. Karkkainen, J. Saastamoinen, M. Tuononen, and P. Franti. Maximum a Posteriori Adaptation of the Centroid Model for Speaker Verification. In *IEEE Signal Processing Letters*, 15, pp. 162–165, 2008.

66. S. Haykin. *Adaptive Filter Theory*. 4th edition. Prentice Hall, 2001.

67. M. Hiipakka. *Measurement Apparatus and Modelling Techniques of Ear Canal Acoustics*. Technical Report, Department of Signal Processing and Acoustics, Helsinki University of Technology, 2008.

68. P. Hofman and J. Van Opstal. Spectro-Temporal Factors in Two-Dimensional Human Sound Localization. In *The Journal of the Acoustical Society of America*, 103(5), pp. 2634–2648, 1998.

69. H. Hu, L. Zhou, H. Ma, and Z. Wu. Head Related Transfer Function Personalization Based on Multiple Regression Analysis. In *Proceedings of the International Conference on Computational Intelligence and Security*, pp. 1829–1832. 2006.

70. M. Hyder, M. Haun, and C. Hoene. Placing the Participants of a Spatial Audio Conference Call. In *Proceedings of the Consumer Communications and Networking Conference-Multimedia Communication and Services*, pp. 1–7. 2010.

71. A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons, 2001.

72. IRCAM, Room Acoustics Team. Listen HRTF Database. URL `http://recherche.ircam.fr/equipes/salles/listen`. Accessed: 19.04.2014.

73. M. Isard and A. Blake. Condensation—Conditional Density Propagation for Visual Tracking. In *International Journal of Computer Vision*, 29(1), pp. 5–28, 1998.

74. K. Ishiguro, T. Yamada, S. Araki, T. Nakatani, and H. Sawada. Probabilistic Speaker Diarization With Bag-of-Words Representations of Speaker Angle Information. In *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2), pp. 447–460, 2012.

75. ITU-R Recommendation BS.1116-1. Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems. International Telecommunication Union, 1997.

76. ITU-T Recommendation E.800. Definitions of Terms Relataed to Quality of Service. International Telecommunication Union, 2008.

77. ITU-T Recommendation G.114. One-Way Transmission Time. International Telecommunication Union, 2003.

78. ITU-T Recommendation P.10/G.100. Vocabulary for Performance and Quality of Service (Amendment 2). International Telecommunication Union, 2008.

79. ITU-T Recommendation P.1301. Subjective Quality Evaluation of Audio and Audiovisual Multiparty Telemeetings. International Telecommunication Union, 2003.

80. ITU-T Recommendation P.800. Methods for Subjective Determination of Transmission Quality. International Telecommunication Union, 1996.

81. Y. Iwaya. Individualization of Head-Related Transfer Functions With Tournament-Style Listening Test: Listening With Other's Ears. In *Acoustical Science and Technology*, 27(6), pp. 340–343, 2006.

82. I. Jolliffe. *Principal Component Analysis*. 2nd edition. Springer, 2002.

83. D. Jones, K. Stanney, and H. Foaud. An Optimized Spatial Audio System for Virtual Training Simulations: Design and Evaluation. In *Proceedings of the International Conference on Auditory Display*, pp. 223–227. 2005.

84. H. Kayser, S. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier. Database of Multichannel In-Ear and Behind-the-Ear Head-Related and Binaural Room Impulse Responses. In *EURASIP Journal on Advances in Signal Processing*, 2009.

85. P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel. Speaker and Session Variability in GMM-Based Speaker Verification. In *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4), pp. 1448–1460, 2007.

86. F. Keyrouz and K. Diepold. Binaural Source Localization and Spatial Audio Reproduction for Telepresence Applications. In *Presence: Teleoperators and Virtual Environments*, 16(5), pp. 509–522, 2007.

87. R. Kilgore, M. Chignell, and P. Smith. Spatialized Audioconferencing: What are the Benefits? In *Proceedings of the Conference of the Centre for Advanced Studies on Collaborative Research*, pp. 135 – 144. 2003.

88. T. Kinnunen and L. Haizhou. An Overview of Text-Independent Speaker Recognition: From Features to Supervectors. In *Speech Communication*, 52(1), pp. 12–40, 2010.

89. D. Kistler and F. Wightman. A Model of Head-Related Transfer Functions Based on Principal Components Analysis and Minimum-Phase Reconstruction. In *The Journal of the Acoustical Society of America*, 91(3), pp. 1637–1647, 1992.

90. C. Knapp and G. Carter. The Generalized Correlation Method for Estimation of Time Delay. In *IEEE Transactions on Acoustics, Speech and Signal Processing*, 24(4), pp. 320–327, 1976.

91. C. Kozielski, M. Rothbucher, and K. Diepold. *Online Speaker Recognition for Teleconferencing Systems*. Technical Report, Institute for Data Processing, Technische Universität München, 2014.

92. A. Kuhn, M. Rothbucher, and K. Diepold. *HRTF Customization by Regression*. Technical Report, Institute for Data Processing, Technische Universität München, 2014.

93. L. Lathauwer, B. Moor, and J. Vandewalle. A Multilinear Singular Value Decomposition. In *SIAM Journal on Matrix Analysis and Applications*, 21(4), pp. 1253–1278, 2000.

94. I. Lee, T. Kim, and T. Lee. Fast Fixed-Point Independent Vector Analysis Algorithms for Convolutive Blind Source Separation. In *Signal Processing*, 87(8), pp. 1859–1871, 2007.

95. I. Lee and T. Lee. On the Assumption of Spherical Symmetry and Sparseness for the Frequency-Domain Speech Model. In *IEEE Transactions on Audio, Speech, and Language Processing*, 15(5), pp. 1521–1528, 2007.

96. Y. Lee, Y. Park, and Y. Park. Newly Designed HRTF Measuring System. In *Proceedings of the International Joint Conference of the Institute of Control, Robotics and Systems and the Society of Instrument and Control*, pp. 1781–1784. 2009.

97. T. Letowski. Sound Quality Assessment: Concepts and Criteria. In *Audio Engineering Society Convention*. 1989.

98. J. Lewald and W. Ehrenstein. Auditory-Visual Spatial Integration: A New Psychophysical Approach Using Laser Pointing to Acoustic Targets. In *The Journal of the Acoustical Society of America*, 104(3), pp. 1586–1597, 1998.

99. C. Liu, B. Wheeler, W. O'Brien Jr, R. Bilger, C. Lansing, and A. Feng. Localization of Multiple Sound Sources With Two Microphones. In *The Journal of the Acoustical Society of America*, 108(4), pp. 1888–1905, 2000.

100. E. Lopez-Poveda and R. Meddis. A Physical Model of Sound Diffraction and Reflections in the Human Concha. In *The Journal of the Acoustical Society of America*, 100(5), pp. 3248 – 3259, 1996.

101. G. Lorho. *Percieved Quality Evaluation - An Application to Sound Reproduction over Headphones*. Ph.D. thesis, Aalto University School of Science and Technology, 2010.

102. R. Lyon. A Computational Model of Binaural Localization and Separation. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 8, pp. 1148–1151. 1983.

103. M. Lüftl, M. Rothbucher, and K. Diepold. *Effizienz und Effektivität eines 3D-Telekonferenz-Systems*. Technical Report, Institute for Data Processing, Technische Universität München, 2014.

104. J. MacDonald. A Localization Algorithm Based on Head-Related Transfer Functions. In *The Journal of the Acoustical Society of America*, 123(6), pp. 4290–4296, 2008.

105. C. Mackersie and H. Cones. Subjective and Psychophysiological Indexes of Listening Effort in a Competing-Talker Task. In *Journal of the American Academy of Audiology*, 22(2), pp. 113–122, 2011.

106. S. Makino, T. Lee, and H. Sawada. *Blind speech separation*. Signals and Communication Technology. Springer, 2007.

107. J. Makous and J. Middlebrooks. Two Dimensional Sound Localization by Human Listeners. In *The Journal of the Acoustical Society of America*, 87(5), pp. 2188–2200, 1990.

108. H. Martens and M. Martens. *Multivariate Analysis of Quality. An Introduction*. John Wiley & Sons, 2001.

109. M. Meier, M. Weitnauer, R. Neudel, and O. Pidancet. *3D VIVANT - Deliverable 3.10, 3D Audio Content Generation*. Technical Report, School of Engineering and Design, Brunel University, 2012.

110. T. Melia and S. Rickard. Underdetermined Blind Source Separation in Echoic Environments Using DESPRIT. In *EURASIP Journal on Applied Signal Processing*, 2007(1), 2007.

111. J. Middlebrooks and D. Green. Observations on a Principal Components Analysis of Head-Related Transfer Functions. In *The Journal of the Acoustical Society of America*, 92(1), pp. 597–599, 1992.

112. P. Minnaar, 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), pp. 323–336, 2001.

113. H. Møller, M. Sørensen, C. Jensen, and D. Hammershøi. Binaural Technique: Do We Need Individual Recordings? In *Journal of the Audio Engineering Society*, 44(6), pp. 451–469, 1996.

114. S. Möller. *Quality of Telephone-Based Spoken Dialogue Systems*. Springer, 2004.

115. S. Möller. *Quality Engineering*. Springer, 2010.

116. H. Møller, M. Sørensen, D. Hammershøi, and C. Jensen. Head-Related Transfer Functions of Human Subjects. In *Journal of the Audio Engineering Society*, 43(5), pp. 300–321, 1995.

117. T. Nishino, N. Inoue, K. Takeda, and F. Itakura. Estimation of HRTFs on the Horizontal Plane Using Physical Features. In *Applied Acoustics*, 68(8), pp. 897–908, 2007.

118. K. Otsuka, S. Araki, K. Ishizuka, M. Fujimoto, M. Heinrich, and J. Yamato. A Realtime Multimodal System for Analyzing Group Meetings by Combining Face Pose Tracking and Speaker Diarization. In *Proceedings of the International Conference on Multimodal Interfaces*, pp. 257–264. 2008.

119. L. Parra and C. Alvino. Geometric Source Separation: Merging Convolutive Source Separation With Geometric Beamforming. In *IEEE Transactions on Speech and Audio Processing*, 10(6), pp. 352–362, 2002.

120. P. Paukner, M. Rothbucher, and K. Diepold. *Sound Localization Performance Comparison of Different HRTF-Individualization Methods*. Technical Report, Institute for Data Processing, Technische Universität München, 2014.

121. M. Pec, M. Bujacz, and P. Strumiłło. Head Related Transfer Functions Measurement and Processing for the Purpose of Creating a Spatial Sound Environment. In *Proceedings of Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments*. 2008.

122. J. Pedersen and T. Jorgensen. *Localization Performance of Real and Virtual Sound Sources*. Technical Report, AM3D, Aalborg Denmark, 2005.

123. T. Plutka, M. Rothbucher, and K. Diepold. *Evaluation of a Channel Assignment Algorithm*. Technical Report, Institute for Data Processing, Technische Universität München, 2014.

124. A. Raake. *Speech Quality of VoIP - Assessment and Prediction*. John Wiley & Sons, 2006.

125. D. Reynolds. Experimental Evaluation of Features for Robust Speaker Identification. In *IEEE Transactions on Speech and Audio Processing*, 2(4), pp. 639–643, 1994.

126. D. Reynolds, T. Quatieri, and R. Dunn. Speaker Verification Using Adapted Gaussian Mixture Models. In *Digital Signal Processing*, 10(1-3), pp. 19–41, 2000.

127. D. Reynolds and R. Rose. Text Independent Speaker Identification Using Automatic Acoustic Segmentation. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, volume 1, pp. 293–296. 1990.

128. S. Rickard and O. Yilmaz. On the Approximate W-Disjoint Orthogonality of Speech. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pp. 529–532. 2002.

129. K. Riederer. Part IIIa: Effect of Microphone Position Changes on Blocked Cavum Conchae Head-Related Transfer Functions. In *Proceedings of the International Congress on Acoustics*, pp. 785–790. 2004.

130. K. Riederer. Part Va: Effect of Head Movement in HRTF Measurements. In *Proceedings of the International Congress on Acoustics*, pp. 795–798. 2004.

131. D. Rife and J. Vanderkooy. Transfer-Function Measurement with Maximum-Length Sequences. In *Journal of the Audio Engineering Society*, 37(6), pp. 419–444, 1989.

132. M. Rothbucher, C. Denk, and K. Diepold. Robotic Gaze Control Using Reinforcement Learning. In *Proceedings of the International Workshop on Haptic Audio Visual Environments and Games*, pp. 83–88. 2012.

133. M. Rothbucher, C. Denk, and K. Diepold. *Robotic Sound Source Separation Using Independent Vector Analysis*. Technical Report, Institute for data processing, Technische Universität München, 2014.

134. M. Rothbucher, M. Durkovic, T. Habigt, H. Shen, and K. Diepold. HRTF-Based Localization and Separation of Multiple Sound Sources. In *IEEE International Symposium on Robot and Human Interactive Communication*, pp. 1092–1096. 2012.

135. M. Rothbucher, M. Durkovic, H. Shen, and K. Diepold. HRTF Customization Using Multiway Array Analysis. In *Proceedings of the European Signal Processing Conference*, pp. 229 – 233. 2010.

136. M. Rothbucher, T. Habigt, J. Feldmaier, and K. Diepold. Integrating HRTF Sound Synthesis into Mumble. In *Proceedings of the International Workshop on Mulitmedia Signal Processing*, pp. 24–28. 2010.

137. M. Rothbucher, T. Habigt, J. Habigt, T. Riedmaier, and K. Diepold. Measuring Anthropometric Data for HRTF Personalization. In *Proceedings of the International Conference on Signal-Image Technology and Internet-Based Systems*, pp. 102 –106. 2010.

138. M. Rothbucher, M. Kaufmann, J. Feldmaier, T. Habigt, M. Durkovic, C. Kozielski, and K. Diepold. 3D Audio Conference System with Backward Compatible Conference Server using HRTF Synthesis. In *Journal of Multimedia Processing and Technologies*, 2(4), pp. 159–175, 2011.

139. M. Rothbucher, D. Kronmüller, H. Shen, and K. Diepold. Underdetermined Binaural 3D Sound Localization Algorithm for Simulaneous Active Sources. In *Audio Engineering Society Convention*. 2010.

140. M. Rothbucher, H. Shen, and K. Diepold. Dimensionality Reduction in HRTF by Using Multiway Array Analysis. In *Human Centered Robot Systems*, pp. 103–110. Springer, 2009.

141. M. Rothbucher, K. Veprek, P. Paukner, T. Habigt, and K. Diepold. Comparison of Head-Related Impulse Response Measurment Approaches. In *The Journal of the Acoustical Society of America*, 134(2), pp. EL223–EL229, 2013.

142. M. Rothbucher, M. Kaufmann, T. Habigt, J. Feldmaier, and K. Diepold. Backwards Compatible 3D Audio Conference Server Using HRTF Synthesis and SIP. In *International IEEE Conference on Signal-Image Technologies and Internet-Based System*, pp. 111–117, 2011.

143. M. Rothbucher, A. Kuhn, and K. Diepold. *HRTF Customization Using the LDV HRTF-database*. Technical Report, Institute for Data Processing, Technische Universität München, 2014.

144. M. Rothbucher, P. Paukner, M. Stimpfl, and K. Diepold. *The TUM-LDV HRTF Database*. Technical Report, Institute for Data Processing, Technische Universität München, 2013.

145. M. Rothbucher, T. Volk, T. Habigt, and K. Diepold. *The LDV Audiolab*. Technical Report, Institute for Data Processing, Technische Universität München, 2013.

146. R. Roy and T. Kailath. ESPRIT-Estimation of Signal Parameters Via Rotational Invariance Techniques. In *IEEE Transactions on Acoustics, Speech and Signal Processing,*, 37(7), pp. 984–995, 1989.

147. F. Rumsey. Spatial Quality Evaluation for Reproduced Sound: Terminology, Meaning, and a Scene-Based Paradigm. In *Journal of the Audio Engineering Society*, 50(9), pp. 651–666, 2002.

148. B. Savas and L. Lim. *Best Multilinear Rank Approximation of Tensors With Quasi-Newton Methods on Grassmannians*. Technical Report, Department of Mathematics, Linkpings University, 2008.

149. J. Scheuing and B. Yang. Disambiguation of TDOA Estimates in Multi-Path Multi-Source Environments. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, volume 4, pp. 837–840. 2006.

150. R. Schmidt. A New Approach to Geometry of Range Difference Location. In *IEEE Transactions on Aerospace and Electronic Systems*, 8(6), pp. 821 – 835, 1972.

151. R. Schmidt. Multiple Emitter Location and Signal Parameter Estimation. In *IEEE Transactions on Antennas and Propagation*, 34(3), pp. 276–280, 1986.

152. M. Schroeder. Integrated Impulse Method Measuring Sound Decay Without Using Impulses. In *The Journal of the Acoustical Society of America*, 66(2), pp. 497–500, 1979.

153. B. Seeber. *Untersuchung der auditiven Lokalisation mit einer Lichtzeigermethode*. Ph.D. thesis, Technische Universität München, 2003.

154. B. Seeber and H. Fastl. Subjective Selection of Non-Individual Head-Related Transfer Functions. In *Proceedings of the International Conference on Auditory Display*, pp. 259 – 262. 2003.

155. C. Segura, A. Abad, J. Hernando, and C. Nadeu. Multispeaker Localization and Tracking in Intelligent Environments. In *Multimodal Technologies for Perception of Humans*, pp. 82–90. Springer, 2008.

156. S. Seok. *Handbook of Research on Human Cognition and Assistive Technology*. Medical Information Science Reference, 2010.

157. J. Skowronek and A. Raake. Investigating the Effect of Number of Interlocutors on the Quality of Experience for Multi-Party Audio Conferencing. In *Proceedings of the Conference of the International Speech Communication Association*, pp. 829–832. 2011.

158. J. Skowronek, A. Raake, K. Hoeldtke, and M. Geier. Speech Recordings for Systematic Assessment of Multi-Party Conferencing. In *Proceedings of Forum Acusticum*, pp. 111–116. 2011.

159. M. Song, C. Zhang, D. Florencio, and H. Kang. Personal 3D Audio System With Loudspeakers. In *Proceedings of the International Conference on Multimedia and Expo*, pp. 1600–1605. 2010.

160. K. Sonntag. *Lehrbuch Arbeitspsychologie.* 3rd edition. Huber, 2012.

161. K. Steierer, M. Rothbucher, and K. Diepold. *Teleconference Channel Assignment*. Technical Report, Institute for Data Processing, Technische Universität München, 2013.

162. S. Takane, D. Arai, T. Miyajima, K. Watanabe, Y. Suzuki, and T. Sone. A Database of Head-Related Transfer Functions in Whole Directions on Upper Hemisphere. In *Acoustic Science and Technology*, 23(3), 2002.

163. Y. Tamai, S. Kagami, Y. Amemiya, Y. Sasaki, H. Mizoguchi, and T. Takano. Circular Microphone Array for Robot's Audition. In *Proceedings of IEEE Sensors*, volume 2, pp. 565–570. 2004.

164. Y. Tamai, S. Kagami, H. Mizoguchi, Y. Amemiya, K. Nagashima, and T. Takano. Realrime 2 dimensional sound source localization by 128-channel huge microphone array. In *Proceedings of the International Workshop on Robot and Human Interactive Communication*, pp. 65–70. 2004.

165. N. Thian, C. Sanderson, and S. Bengio. Spectral Subband Centroids as Complementary Features for Speaker Authentication. In *Biometric Authentication*, pp. 631–639. Springer, 2004.

166. S. Tranter and D. Reynolds. An Overview of Automatic Speaker Diarization Systems. In *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5), pp. 1557–1565, 2006.

167. G. Triebig. *Arbeitsmedizin*. 2nd edition. Gentner, 2008.

168. M. Unvervdorben, M. Rothbucher, and K. Diepold. *Blind Source Separation for Speaker Recognition Systems*. Technical Report, Institute for Data Processing, Technische Universität München, 2014.

169. M. Usman, F. Keyrouz, and K. Diepold. Real Time Humanoid Sound Source Localization and Tracking in a Highly Reverberant Environment. In *Proceedings of the International Conference on Signal Processing*, pp. 2661–2664. 2008.

170. J. Valin, F. Michaud, and J. Rouat. Robust Localization and Tracking of Simultaneous Moving Sound Sources Using Beamforming and Particle Filtering. In *Robotics and Autonomous Systems*, 55(3), pp. 216–228, 2007.

171. J. Valin, F. Michaud, J. Rouat, and D. Letourneau. Robust Sound Source Localization Using a Microphone Array on a Mobile Robot. In *Proceedings of the International Conference on Intelligent Robots and Systems*, volume 2, pp. 1228–1233. 2003.

172. J. Valin, J. Rouat, and F. Michaud. Enhanced Robot Audition Based on Microphone Array Source Separation with Post-Filter. In *Proceedings of the International Conference on Intelligent Robots and Systems*, volume 3, pp. 2123–2128. 2004.

173. B. Van Veen and K. Buckley. Beamforming: A Versatile Approach to Spatial Filtering. In *IEEE Magazine on Acoustics, Speech, and Signal Processing*, 5(2), pp. 4–24, 1988.

174. E. Vincent, R. Gribonval, and C. Févotte. Performance Measurement in Blind Audio Source Separation. In *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4), pp. 1462–1469, 2006.

175. E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. Rosca. First Stereo Audio Source Separation Evaluation Campaign: Data, Algorithms and Results. In *Independent Component Analysis and Signal Separation*, pp. 552–559. Springer, 2007.

176. T. Volk, M. Rothbucher, and K. Diepold. *Quality of Experience - Evaluierung eines Telekonferenzsystems in der Entwicklungsphase*. Technical Report, Institute for data processing, Technische Universität München, 2013.

177. D. Wang and G. Brown. *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley-IEEE Press, 2006.

178. J. Wang, T. Kuan, J. Wang, and G. Gu. Ubiquitous and Robust Text-Independent Speaker Recognition for Home Automation Digital Life. In *Ubiquitous Intelligence and Computing*, pp. 297–310. Springer, 2008.

179. T. Weissgerber, K. Laumann, G. Theile, and H. Fastl. Headphone Reproduction via Loudspeakers using Inverse HRTF Filters. In *Proceedings of the Annual German Congress on Acoustics*, volume 1, pp. 1291–1294. 2009.

180. S. Wold, M. Sjöström, and L. Eriksson. PLS-Regression: A Basic Tool of Chemometrics. In *Chemometrics and Intelligent Laboratory Systems*, 58(2), pp. 109–130, 2001.

181. C. Wooters and M. Huijbregts. The ICSI RT07s Speaker Diarization System. In *Multimodal Technologies for Perception of Humans*, pp. 509–519. Springer, 2008.

182. J. Yang, D. Zhang, A. Frangi, and J. Yang. Two-Dimensional PCA: A New Approach to Appearance-Based Face Representation and Recognition. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(1), pp. 131–137, 2004.

183. J. Ye. Generalized Low Rank Approximations of Matrices. In *Machine Learning*, 61(1), pp. 167–191, 2005.

184. X. Zha, H. Fuchs, and M. Späh. Ein neues Konzept für akustische Freifeldräume. In *Rundfunktechnische Mitteilungen*, 42(3), pp. 81–91, 1998.

185. X. Zha, M. Späh, H. Fuchs, G. Eckholdt, and G. Babuke. Neuer Reflexionsarmer Raum für den gesamten Hörbereich. In *Fraunhofer Institut für Bauphysik: IBP-Mitteilung*, 24(321), 1997.

186. N. Zheng, T. Lee, and P. Ching. Integration of Complementary Acoustic Features for Speaker Recognition. In *IEEE Signal Processing Letters*, 14(3), pp. 181–184, 2007.

187. S. Zilienski, F. Rumsey, and S. Bech. On Some Biases Encountered in Modern Audio Quality Listening Tests - A Review. In *Journal of the Audio Engineering Society*, 56(6), p. 427–451, 2008.