Aachener Beiträge zur Hörtechnik und Akustik

Michael Kohnen

Combining Reproduction Techniques for Room Auralizations

IHTA Institute for Hearing Technology and Acoustics
RWTH AACHEN UNIVERSITY

Combining reproduction techniques for room auralizations

Von der Fakultät für Elektrotechnik und Informationstechnik der
Rheinisch-Westfälischen Technischen Hochschule Aachen
zur Erlangung des akademischen Grades eines

DOKTORS DER INGENIEURWISSENSCHAFTEN

genehmigte Dissertation

vorgelegt von

Master of Science
Michael Kohnen

aus Bocholt, Deutschland

Berichter:

Univ.-Professor Dr. rer. nat. Michael Vorländer
Professor Dr.-Ing. Bruno Sanches Masiero

Tag der mündlichen Prüfung: 5. November 2024

Abstract

3-D sound is an important feature of human hearing and the realistic reproduction of 3-D sound, a key requirement for reliable studies on human behavior and well-being in complex acoustic scenes. Loudspeaker-based sound reproduction provides low-frequency effects on the whole body and lowers the inhibition threshold since no devices need to be attached to the user. Yet, none of the known methods of loudspeaker-based reproduction can provide a good sensation of all perceptual qualities (e.g., good source localization or low coloration) and all of them are influenced by the room they are placed in. To increase the realism of loudspeaker-based auralizations, this thesis develops and evaluates a hybrid system that combines the different strengths of each method. Furthermore, the influence of the room in which the loudspeakers are placed is mitigated by leveraging knowledge of the virtual room and its room impulse response (RIR).

For a seamless transition between the reproduction methods, two quantitative listening experiments are shown that evaluate the loudness differences between the methods and the provided localization accuracy of the individual methods. The perceptual qualities between the methods were compared in a comprehensive listening experiment to find suitable reproduction methods for each time section of the RIR using the Spatial Audio Quality Inventory (SAQI). Finally, a hybrid system is presented that utilizes Crosstalk Cancellation (CTC) for the direct sound, Vector-Base Amplitude Panning (VBAP) for early reflections and Higher Order Ambisonics (HOA) for the diffuse decay. To suppress unwanted early reflections in the reproduction room, the RIR of the virtual room is altered to consider the RIR of each loudspeaker in the reproduction room and evaluated by a listening experiment. For the diffuse decay, the absorption coefficients in the virtual room are adapted so that the energy decay of the virtual and repro-

duction room combined matches the intended room. In a second SAQI listening experiment, the hybrid system is evaluated against a pure CTC system, including an evaluation of the listening room compensation and the provided localization accuracy. Lastly, the integration of the user into the virtual scene is evaluated in a listening experiment with participants speaking into the room.

The results of the loudness experiment show variances around the just noticeable difference of 1 dB within each reproduction method and between methods. The CTC system provides better localization accuracy than VBAP in the median plane, with variations comparable to the natural localization blur found in human hearing. HOA and VBAP show almost no differences in perceptual qualities and show the same differences towards the CTC system in terms of better perception of clarity, naturalness, speech intelligibility, crispness and presence, while at the same time reducing comb filter effects and metallic tone color. The listening room compensation for early reflections provides source perception with less apparent source width in a headphone experiment. The SAQI comparison showed that the hybrid system benefits from the improved perceptual qualities of VBAP and HOA and, unexpectedly, increasing the localization accuracy compared to the CTC system. The effect of the two room compensation approaches, however, were not audible, likely due to the acoustically optimized listening room. A listening experiment regarding user integration showed that the system can present reflections in real-time and integrate the user into the virtual room.

The proposed hybrid system effectively demonstrates the feasibility and advantages of combining reproduction methods. It enhances perceptual quality compared to a pure CTC system and even improves localization accuracy, particularly in cases involving real-time binaural synthesis.

The current implementation of the system is quite complex, especially when it comes to the compensation of early reflections, which presents significant demands. However, it is likely that using just two distinct reproduction methods could be adequate, simplifying the system while still achieving satisfactory results. Additionally, the need for the listening room compensation is less critical if the listening room is already acoustically treated. In such cases, omitting the compensation would help lower the system's requirements without significantly affecting performance.

Kurzfassung

Dreidimensionale Schallwahrnehmung ist das wichtigste Merkmal menschlichen Hörens und die realistische Wiedergabe eine Grundvoraussetzung, um menschliches Verhalten und Wohlbefinden in komplexen akustischen Szenen zu studieren. Die lautsprecherbasierte Wiedergabe ermöglicht die tieffrequente Schallwahrnehmung über den gesamten Körper und hebt die Akzeptanz für die Technologie, da keine Geräte am Benutzer angebracht werden müssen. Keine der bekannten lautsprecherbasierten Wiedergabemethoden ist jedoch in der Lage, alle Qualitäten der Wahrnehmung (wie z. B. gute Quellenlokalisierung oder geringe Klangfärbung) zufriedenstellend zu vermitteln. Zudem werden alle Verfahren von dem Raum beeinflusst, in dem sie platziert sind. Um den Realismus der Wiedergabe zu erhöhen, wurde in dieser Arbeit ein hybrides System entwickelt und evaluiert, das die unterschiedlichen Stärken der einzelnen Methoden kombiniert und den Einfluss des Raums, in dem die Lautsprecher aufgestellt sind, unterdrückt. Dazu werden Informationen über den virtuellen Raum und seiner Raumimpulsantwort (RIR) genutzt.

Für einen nahtlosen Übergang zwischen den Wiedergabemethoden werden zwei Hörexperimente vorgestellt, welche die Lautstärkeunterschiede und Lokalisationsgenauigkeit zwischen den Methoden quantifizieren. Die Wahrnehmungsqualitäten zwischen den Methoden wurden in einem umfassenden Hörexperiment verglichen, um geeignete Wiedergabemethoden für jeden Zeitabschnitt der RIR mithilfe des Spatial Audio Quality Inventory (SAQI) zu finden. Abschließend wird ein hybrides System vorgestellt, das Crosstalk Cancellation (CTC) für den Direktschall, Vector-Base Amplitude Panning (VBAP) für frühe Reflexionen und Higher Order Ambisonics (HOA) für den diffusen Nachhall verwendet. Um unerwünschte frühe Reflexionen im Wiedergaberaum zu unterdrücken, wird die RIR des virtuellen Raums so verändert, dass die RIRs aller Lautsprecher im Wiedergaberaum berücksichtigt werden und durch ein Hörexperiment evaluiert. Für den diffusen Nachhall werden die Absorptionskoeffizienten im virtuellen Raum

so angepasst, dass der Energieabfall des Virtuellen und des Wiedergaberaums zusammen dem Zielraum entspricht. In einem zweiten SAQI Hörexperiment wird das hybride System einem reinen CTC System gegenübergestellt und zusätzlich die bereitgestellte Lokalisationsgenauigkeit verglichen sowie die Hörraumkompensation evaluiert. Abschließend wird die Integration des Benutzers in die virtuelle Szene in einem Hörexperiment, in dem die Teilnehmer in einen virtuellen Raum sprechen, evaluiert.

Die Ergebnisse des Lautstärkeexperiments zeigen Abweichungen von etwa 1 dB innerhalb der einzelnen Wiedergabemethoden und zwischen den Methoden. Das CTC System bietet eine bessere Lokalisationsgenauigkeit als VBAP in der Medianebene mit Abweichungen, die mit der natürlichen Lokalisationsunschärfe des menschlichen Gehörs vergleichbar sind. HOA und VBAP zeigen fast keine Unterschiede in den Wahrnehmungsqualitäten und weisen die gleichen Unterschiede zur CTC in Bezug auf eine bessere Wahrnehmung von Klarheit, Natürlichkeit, Sprachverständlichkeit, Schärfe und Präsenz auf, während gleichzeitig die Wahrnehmung von Kammfiltereffekten und metallischer Klangfarbe reduziert werden. In einem Kopfhörerversuch wird zudem gezeigt, dass die Hörraumkompensation für frühe Reflexionen die Wahrnehmung der Quellenbreite verringert. Das zweite SAQI Experiment zeigt, dass das hybride System von den verbesserten Wahrnehmungsqualitäten von VBAP und HOA profitiert und unerwarteterweise die Lokalisationsgenauigkeit im Vergleich zur CTC erhöht. Die Effekte der beiden Raumkompensationsansätze waren nicht hörbar, was wahrscheinlich an dem akustisch optimierten Hörraum liegt. Ein Hörexperiment zur Benutzerintegration zeigt, dass das System in der Lage ist, Reflexionen in Echtzeit darzustellen und den Benutzer in den virtuellen Raum zu integrieren.

Das vorgestellte hybride System demonstriert effektiv die Machbarkeit und Vorteile der Kombination von Wiedergabemethoden. Es verbessert die Wahrnehmungsqualität im Vergleich zur reinen CTC und verbessert sogar die Lokalisationsgenauigkeit, insbesondere bei binauraler Synthese in Echtzeit.

Die derzeitige Implementierung des Systems ist komplex und die Kompensation früher Reflexionen stellt erhebliche Anforderungen an die Modellierung des Hörraums. Es ist jedoch wahrscheinlich, dass die Verwendung von nur zwei unterschiedlichen Wiedergabemethoden ausreichend ist, um das System zu vereinfachen. Weiterhin kann die Hörraumkompensation vernachlässigt werden, wenn der Hörraum bereits akustisch optimiert ist und so die Systemanforderungen weiter senken, ohne das Ergebnis signifikant zu beeinträchtigen.

Contents

Glossary xi

1 Introduction 1
- 1.1 The idea of a hybrid system . . . 4
- 1.2 Research question . . . 5

2 Fundamentals of sound presentation 7
- 2.1 Coordinate system and object related angles . . . 8
- 2.2 Spatial hearing . . . 9
  - 2.2.1 Cone-of-Confusion . . . 11
- 2.3 Spatial Audio Quality Inventory . . . 11
- 2.4 Perception of room impulse responses . . . 12
  - 2.4.1 Direct sound . . . 12
  - 2.4.2 Early reflections . . . 13
  - 2.4.3 Diffuse decay . . . 13
  - 2.4.4 Mixing times . . . 14
- 2.5 Rendering and room simulation . . . 15
- 2.6 CTC - Crosstalk cancellation . . . 16
  - 2.6.1 Perceptual aspects . . . 19
- 2.7 VBAP - Vector-Base Amplitude Panning . . . 19
  - 2.7.1 Perceptual aspects . . . 21
- 2.8 HOA - Higher Order Ambisonics . . . 22
  - 2.8.1 Perceptual aspects . . . 24

3 Setup and implementation 25
- 3.1 Virtual Reality Laboratory . . . 26

3.2 Loudspeaker array . . . 27
3.3 Artificial head and HRTF . . . 28
3.4 Crosstalk cancellation . . . 29
3.5 Vector-Base Amplitude Panning . . . 29
3.6 Higher Order Ambisonics . . . 30
3.7 Listening tests . . . 30
3.7.1 Positioning of participants . . . 30
3.7.2 Sound source positions . . . 31
3.7.3 Virtual rooms . . . 33

4 Perceptual qualities of reproduction systems 35
4.1 Loudness . . . 36
4.1.1 Method . . . 37
4.1.2 Setup . . . 37
4.1.3 Results . . . 38
4.1.4 Discussion . . . 40
4.1.5 Conclusion . . . 41
4.2 Localization . . . 41
4.2.1 Method . . . 42
4.2.2 Setup . . . 42
4.2.3 Results . . . 43
4.2.4 Discussion . . . 45
4.2.5 Conclusion . . . 46
4.3 Spatial Audio Quality . . . 46
4.3.1 Method . . . 47
4.3.2 Setup . . . 47
4.3.3 Results . . . 52
4.3.4 Discussion . . . 55
4.3.5 Conclusion . . . 56

5 Compensation of the listening room 59
5.1 Early reflections . . . 60
5.1.1 Listening room - hybrid model . . . 61
5.1.2 Reflection detection . . . 62
5.1.3 Reflection distance . . . 62
5.1.4 Implementation . . . 63
5.1.5 Subjective proof-of-concept . . . 64
5.2 Late reverberation . . . 68

6 Hybrid system **71**
6.1 Implementation 72
6.1.1 Loudness 74
6.1.2 Latency 75
6.2 Subjective evaluation 78
6.2.1 Spatial Audio Quality Inventory 79
6.2.2 User integration 87

7 Conclusion **95**

8 Summary and outlook **99**
8.1 Summary 100
8.2 Outlook 103

Acknowledgements **105**

Curriculum Vitae **107**

A Appendix **109**

Bibliography **115**

Glossary

Notation

$\overrightarrow{\mathbf{g}}$	Spatial vector
$\mathbf{G}$	Matrix
$\mathbf{g}$	Matrix vector

Mathematical operators

$(\cdot)^*$	Hermitian transpose
$(\cdot)^\dagger$	pseudo inverse of a matrix
$\|\cdot\|_2$	L2 norm of a vector
$\log(\cdot)$	common logarithm (base 10)
$(\cdot)^{-1}$	matrix inverse

Symbols

S	Surface area inside a room	m^2
V	Volume of a room	m^3
σ	Standard deviation of a sample	
θ	Elevation angle from horizontal plane, see 2.1	$^\circ$
φ	Azimuth angle in horizontal plane, see 2.1	$^\circ$
c	Speed of sound, approx. 343 m/s	m/s
f	Frequency	Hz
r	Radial distance	m
t_m	Time transitioning to DD after the DS	s

Acronyms

ANOVA	Analysis of Variance
ASW	apparent source width
BRAS	Benchmark for Room Acoustical Simulation, see [Bri+21; Asp+20]
BRIR	binaural room impulse response
CTC	crosstalk cancellation
CVH	hybrid system combining CTC, VBAP and HOA
DD	diffuse decay
DS	direct sound
ER	early reflection
FIR	finite impulse response
GUI	graphical user interface
HMD	head-mounted display
HOA	Higher Order Ambisonics
HpTF	headphone transfer function
HRIR	head-related impulse response
HRTF	head-related transfer function
IHTA	Institute for Hearing Technology and Acoustics, RWTH Aachen University, Germany
ILD	interaural level difference
IR	impulse response
ITD	interaural time difference
JND	just noticeable difference
QUEST	Bayesian adaptive psychometric method [WP83]
RAVEN	Room Acoustics for Virtual Environments [SV11; Sch11]
RIR	room impulse response
RMS	root mean square
RT	reverberation time
SAQI	Spatial Audio Quality Inventory, see Section 2.3
SH	spherical harmonics
significance level	significance level: * for $p<0.05$, ** for $p<0.01$ and *** for $p<0.001$
SPL	sound pressure level
VA	Virtual Acoustics, see [Ins]
VBAP	Vector-Base Amplitude Panning, see Section 2.7
VR	virtual reality
VR-Lab	Virtual Reality Laboratory at IHTA, see Section 3.1

1

Introduction

Our everyday life is filled with the perception of sound. From the enjoyment of podcasts, music, entertainment, or a friendly voice, sound provides us with information about one's emotional state or the calmness of the surrounding environment. While this calmness can be recreational, noise, on the other hand, can affect our health, decrease the quality of communication, lower our cognitive skills, and mask vital information. The recreation of sound allows us to be part of a distant scenery or to relive past moments in our memory. Spatial sound provides us with information about source localization in all three dimensions and, specifically, outside our head. As a result, listening to spatial sound enables us to separate sources from each other, put them in perspective, and focus on each of them individually. This extra dimension provides us with a feeling of "presence", of being there. It allows us to immerse ourselves into a scene and be surrounded by sound. Focusing on single sources allows us to have a conversation with one person, even though many people in the same room are speaking. Spatial sound provides us with vital information about objects moving around us, especially in road traffic, and allows for orientation in complex scenes.

Applications using spatial sound are numerous. Auralizations of virtual scenes can recreate historical sites and events, and extended reality can steer our attention towards events outside the user's field of view intuitively. Similarly, medical equipment uses spatial sound as an efficient warning mechanism. Multi-user communication systems can lower the listening effort by spatially separating the speaker and making musical instruments distinguishable in a mix. Cinemas provide spatial sound for an increased experience of immersion in entertainment. With the rise of head-mounted displays and 3-D visual cameras, the desire for full three-dimensional spatial sound is growing. The presentation of spatial sound can further be used to conduct controlled studies and training. To quantify the effects of noise and disturbing sound sources, controlled sound scenes that recreate the intended sound field as accurately as possible have to be provided. In the same way, training can be applied, for example, to hearing-impaired people.

Various methods are available to present spatial sound. Using headphones, binaural signals can provide spatial audio. Widely known loudspeaker-based methods include the surround sound techniques used in (home) cinema applications, such as Dolby Atmos, DTS, or the 5.1 systems. Other, more scientifically known methods include Crosstalk Cancellation (also known as transaural), Vector-Base Amplitude Panning, and Ambisonics. Finally, different methods for recreating sound fields with a high number of loudspeakers are available, with Wave Field Synthesis being the best known. However, not all of them offer a full three-dimensional presentation of sound. Each of these techniques has different strengths and weak-

nesses.

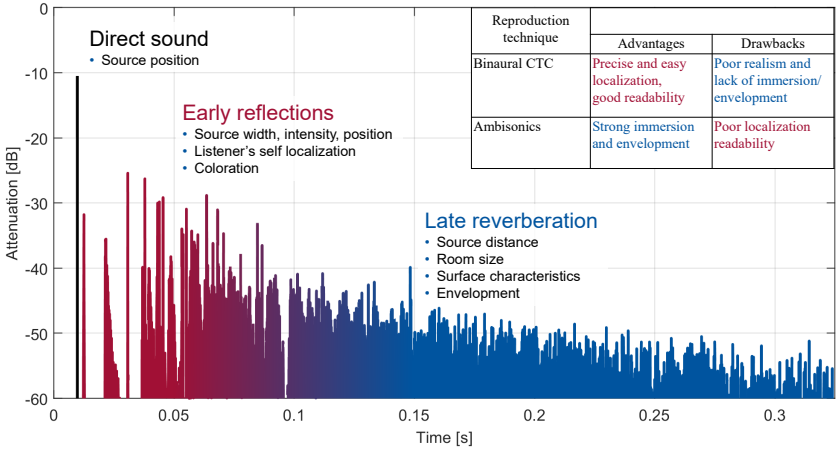
When comparing the techniques, the question emerges: What defines good spatial sound reproduction? Is it the provided accuracy of source positions? Is it the less distorted (also referred to as "colored") sound? Spatial sound systems are defined by several qualities. Comparisons can estimate the differences in these qualities. Based on the application, a suitable technique can be selected. While virtual reality with a focus on interaction demands accurate source locations, the presentation of music prioritizes minimal sound coloration. The selection of suitable terminology, its relevance for spatial audio, and a uniform understanding among people are crucial for assessing meaningful results.

To present spatial audio, a three-dimensional (acoustic) scene is needed. Video or audio recordings of three-dimensional environments are usually bound to one specific position, prohibiting any movement except rotation. Interactive environments, where the user is free to move, are typically synthesized by creating a computer model of the surroundings (e.g., a room), which includes information about sources and the user as a listener. Simulations are run to define the temporal behavior of the sound arriving at the listener in the form of a room impulse response. The arriving sound can be separated into a direct sound (sound traveling along the line of sight), the early reflections (reflections up to a certain time threshold after the direct sound), and a late reverberant part (also called diffuse decay). Each time slot provides certain cues for the overall perception of the virtual scene.

This thesis analyses and compares three loudspeaker-based spatial audio reproduction methods that provide source locations in every direction, to identify their strengths and weaknesses. These results are used to verify their suitability for reproducing the specific perceptual qualities for each of the time sections of the room impulse response. The idea is to find a perfectly matching reproduction system for each time section of the room impulse response. Ideally, this approach suppresses the weaknesses of each reproduction system without compromising their strengths, as suggested in [PSV11; PSV14]. To allow for an inaudible transition between the reproduction systems, loudness and latency adaptations are needed and are investigated. Additionally, when using loudspeaker-based reproduction, reflections of the room in which the loudspeakers are placed are added during playback. Two approaches to compensate these reflections are developed: one for the early reflections and one for the diffuse decay. Finally, the hybrid

system is evaluated to determine whether the desired effects are audible. Furthermore, the integration of the user into the virtual scene is tested.

1.1 The idea of a hybrid system



Reproduction technique	Advantages	Drawbacks
Binaural CTC	Precise and easy localization, good readability	Poor realism and lack of immersion/ envelopment
Ambisonics	Strong immersion and envelopment	Poor localization readability

Figure 1.1: A room impulse response (created by simulation) separated into the time sections of direct sound, early reflections and diffuse decay. The sections of early reflections and diffuse decay overlap, and the linear color change indicates the transition between these two. The table in the upper right is taken from Guastavino et al. [Gua+07] and shows the potential benefit of combining the reproduction techniques.

The idea of combining different reproduction methods for the auralization of virtual rooms using the presented approach was first formulated by Pelzer et al. [PSV11; PSV14] as an enhancement of the work presented by Favrot and Buchholz [FB10]. They suggest changing the reproduction method due to different events (or time sections) in a room impulse response (RIR). Each time section of the room impulse response provides different perceptual qualities for the listener. Therefore, a reproduction method for a time section only has to provide good cues that are demanded in this specific time section, yet, does not have to provide a sense of perceptual qualities that are demanded by other time sections. The feasibility of this idea is supported by findings of Guastavino et al. [Gua+07] which compared different reproduction methods to each other. Guastavino et al. showed that the crosstalk cancellation crosstalk cancellation (CTC) system provided precise source localization, but "lack of immersion/envelopment". The

direct sound (DS) is mainly important for source localization and would therefore benefit from the CTC system. The lack of envelopment, on the other hand, is assumed to have less effect, as "immersion/envelopment" is provided by the later part of the RIR. Ambisonics, on the other hand, shows the direct opposite of qualities compared to the CTC system: poor localization and "strong immersion and envelopment". Ambisonics would therefore be suitable for the later part of the RIR. Figure 1.1 illustrates the relations. The combination of reproduction techniques would therefore benefit from the strengths of each technique, while at the same time weaknesses are compensated.

In general, a RIR can be separated into three different time sections. The DS, arrival of distinct early reflections (ERs) and a diffuse decay (DD) process. While the separation between DS and ERs is defined, the separation between the ERs and DD is not. The so-called mixing time describes the time between the arrival of the DS and the transition of the sound field into a perceptually diffuse sound field, i.e., position and orientation independent perception of the reverberation. Lindau et al. [LKW12] compared different mixing time estimators and concluded that "on average two orders of reflection, and a reflection density of less than 200 s^{-1} is perceived as diffuse even by trained listeners." The RIR in a geometrical acoustic simulation is typically calculated using a hybrid approach image source model and Ray Tracing, with typical image source orders of two or three, which matches the mixing time prediction by Lindau et al. .

1.2 Research question

The aim of this thesis is to develop a novel reproduction system that utilizes different methods for different sections of a (simulated) RIR that acoustically integrates the user into the virtual scene. To detach the user, and the virtual scene, from the room acoustics present in the listening room (i.e., the room the loudspeaker array is placed in), new types of room compensating measures are developed.

The idea of combing reproduction methods is based on findings using a horizontal loudspeaker ring. In general, findings in literature rely on different set-ups (especially loudspeaker arrays) and different implementations of the reproduction methods. They further use different perceptual qualifiers and testing methods. This thesis will therefore provide a direct comparison of the reproduction methods in the same set-up with the implemented reproduction methods using the Spatial Audio Quality Inventory (SAQI) method in Section 4.3. The comparison

will answer whether the assumptions made are valid in the given context and which reproduction methods might be suitable for the different time sections of the RIR. The provided localization accuracy of a reproduction method is one main feature of spatial audio. In Section 4.2 the provided accuracy of the reproduction systems were tested to select an appropriate system for reproducing the DS. To keep the original structure of the RIR the reproduction methods have to match in loudness. A listening experiment was conducted in Section 4.1 to answer whether loudness adaptation towards a reference source is necessary and to quantify the homogeneity of loudness perception within each reproduction system.

The proposed hybrid reproduction system relies on information about the RIR in the virtual scene with directional information, i.e., the direction of incoming reflections (otherwise a spatial reproduction would not be possible). Chapter 5 investigates the possibility of using this detailed information about the virtual room and combine it with detailed acoustical information about the listening room to compensate the influence of the latter by modifying the RIR of the virtual scene.

Chapter 6 combines the information gathered in the previous chapters to implement the hybrid system. In Section 6.2.1 a listening experiment is made to evaluate the hybrid system, answering the question if improvements are audible and if other perceptual qualities degrade. The experiment further tests the audibility of the proposed room compensation and the provided localization accuracy against a reference. Section 6.1.2 investigates the integration of the user into the virtual scene by latency measurements and perceptually relevant update rates of the ERs to answer whether a real-time response of the system is feasible and to optimize the computational effort for the integration.

2

Fundamentals of sound presentation

Recreating plausible or even authentic sound perception requires different stages of processing, from rendering the virtual scene to calculating the loudspeaker signals. Additionally, understanding the physical sound propagation as well as human sound perception is fundamental. In the rendering and simulation stage, information describing the virtual scene is gathered, and sound propagation from a virtual source to a virtual listener is calculated. In the second stage, the calculated information of the sound field at the virtual listener has to be delivered to the actual listener as accurately as possible. To evaluate the performance of the overall sound reproduction, a vocabulary catalog and testing method called SAQI is introduced.

2.1 Coordinate system and object related angles

For Cartesian coordinates, this thesis uses a right-handed coordinate system similar to the OpenGL conventions. Regarding a human listener, the positive x-axis denotes the right direction, the y-axis the upwards direction, and the z-axis the rear direction. Therefore, the negative z-axis denotes the forward direction. The angular coordinate conventions are shown in Figure 2.1, where the azimuth angle φ is the angle in the horizontal plane from the forward direction towards the left side for angles between 0° and 180°, and to the right side for angles between 0° and -180°. The elevation angle θ describes the angle from the horizontal plane upwards with angles between 0° and 90° and downwards for angles between 0° and -90°. Typical directions can then be described as shown in Table 2.1.

Direction	$\varphi[°]$	$\theta[°]$	x	y	z
Frontal:	0°	0°	0	0	-1
Rear:	180°	0°	0	0	1
Left:	90°	0°	-1	0	0
Right:	270°	0°	1	0	0
Up:	0°	90°	0	1	0
Down:	0°	−90°	0	-1	0

Table 2.1: Typical directions and their coordinates as used in this thesis.

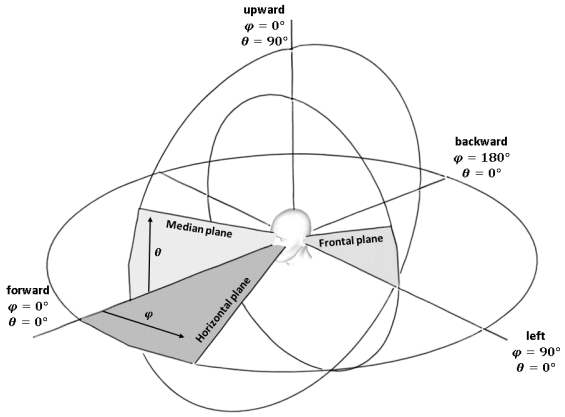


Figure 2.1: Definitions used in this thesis for the coordinate system. The different planes are plotted with different radii to better distinguish them from each other. The azimuthal angle φ increases to the left side, the elevation angle θ increases towards the upward direction. The angular conventions comply with those in [Bla96].

2.2 Spatial hearing

Humans can perceive sound around them and outside their head. They can perceive incoming directions of sound and localize sound sources up to a certain accuracy. To do so, the human auditory system relies on three main cues [Bla96]. The first two rely on differences between the left and right ear and are therefore named binaural differences or binaural cues. For higher frequencies, the human head functions as an acoustic obstacle and, depending on the relative position of the source to the human, results in an amplitude difference between left and right ear, the so-called interaural level difference (ILD). For lower frequencies, the head geometry is small compared to the wavelength and the shadowing effect of the head decreases. Consequently, for lower frequencies, the difference in time of arrival and phase of the sound wave between the left and right ear is more relevant, the so-called interaural time difference (ITD). The transition from evaluating ILD information to a focus on ITD is somewhere around 1500 Hz. The binaural differences change only little in the median plane, where monaural cues dominate the localization mechanisms of human sound perception. Monaural cues are frequency dependent changes of the sound due to reflection and diffraction from the human body, especially those of the ears, and change depending on the

direction of sound incidence and frequency of the sound wave. As the monaural cues rely on the body shape, they are highly individual.

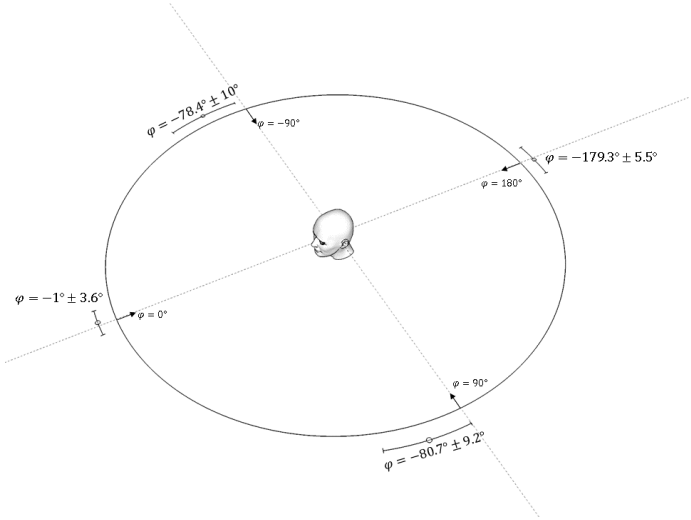


Figure 2.2: Mismatch between the perceived source location and the actual source location for broadband, pulsed noise signals in the horizontal plane when the listener's head is fixed to the frontal direction. Participants had to move a loudspeaker into either frontal, left, rear or right direction. The perceived source direction is indicated by arrows. Participant's response, the actual direction of the incoming source, is indicated outside the ring. The picture is similar to [Bla96].

Blauert [Bla96] summarized findings on human localization accuracy. Figures 2.2 and 2.3 illustrate the essentials of these findings. Figure 2.2 shows the mismatch between the perceived source location and the actual source location for broadband, pulsed noise signals in the horizontal plane when the listener's head is fixed to the frontal direction. For the lateral directions a mismatch close to 10° can be observed as well as a just noticeable difference (JND) of 10° when shifting the real source position. Stimulus content, length of the stimulus, as well as head movement, can decrease the blur and increase the localization accuracy. Figure 2.3 shows results for source localization accuracy in the median plane where binaural cues are close to zero. As input signal, familiar speech was used and the head was fixed. Note that due to the experimental design, the indication of real source and perceived source location is changed between Figure 2.2 and Figure 2.3. For frontal elevated sources the localization blur is found to be ±10°.

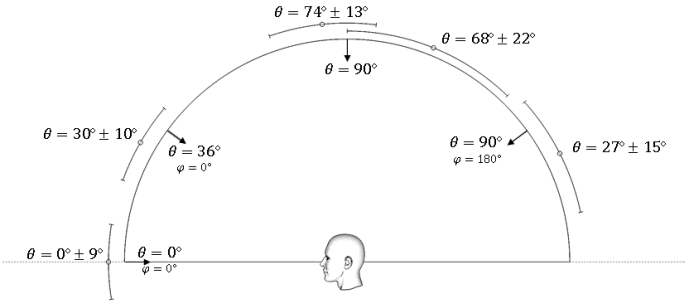


Figure 2.3: Mismatch between the perceived source location and the actual source location for broadband, pulsed noise signals in the median plane for familiar speech when the listener's head is fixed to the frontal direction. Arrows indicate the direction of the sound source, the outer ring the perceived source position. The picture is similar to [Bla96].

The same influencing parameters on localization accuracy as for the horizontal plane can be found.

2.2.1 Cone-of-Confusion

The main cues for horizontal source localization are ILD and ITD. Source positions on the median plane result in almost no time or level difference between the ears. Outside the median plane, the source may travel along a circle shape in a plane parallel to the median plane without changing its ITD or ILD value. For fixed ILD and ITD values, the radius of the circle decreases when moving the parallel plane closer to the median plane and increases when moving the plane farther away. The overall shape of fixed binaural cues then becomes cone-like. The monaural cues are relevant to distinguish different source positions on this cone and, especially when using non-individual head-related transfer functions (HRTFs), front-back confusion may occur on this cone-of-confusion. It should be noted that the human head is not perfectly symmetric, and ILD and ITD are frequency-dependent. Nevertheless, the concept of the cone-of-confusion is a good approximation for understanding the phenomena of front-back confusion.

2.3 Spatial Audio Quality Inventory

The quality of sound presentation covers various aspects, and the importance of each aspect might differ depending on the system's purpose. While virtual reality applications typically prioritize sound localization over uncolored sound,

the emphasis can be the opposite for pure music reproduction, where no visual cues are provided, and the absolute position of e.g., instruments is less important. Therefore, a catalog of perceptual qualities is necessary to evaluate a sound reproduction system comprehensively. The challenging task of accessing spatial audio quality features separates into three main aspects: the completeness of all relevant features, the linguistic clarity and mutual understanding of these features among subjects, and the actual assessment of these features in a test procedure. While different catalogs and procedures exist that focus on evaluating different codecs or algorithms, they usually do not concentrate on altering the reproduction technique and assessing aspects of spatial audio. The SAQI provides a scientific approach to address these three challenges [Lin+14] and to make studies more comparable. 21 spatial audio experts from the scientific field selected relevant perceptual qualities and defined descriptions in a focus group panel, forming a consensus in the (German) scientific field. The test procedure for accessing the perceptual qualities can be found in [Lin14] and involves comparing a stimulus against an inner or given reference. Each comparison begins with whether an overall difference can be heard. If so, the perceptual qualities will be rated on a scale, with the type of scale and terminology for the upper and lower descriptor of the scale being provided by the test manual. One crucial aspect of the test procedure is ensuring that participants properly understand the descriptions given for each perceptual quality. Three audio samples are provided for three out of the 48 different perceptual qualities.

2.4 Perception of room impulse responses

A RIR is the impulse response of a source (as omnidirectional as possible) to a receiver (typically an omnidirectional microphone). While different challenges arise for real measurements, the RIRs in this thesis are gathered by simulation, as described in Section 2.5. Other types of impulse responses, like the binaural room impulse response (BRIR), are direction-dependent. Consequently, the RIR is the universal impulse response discussed in literature. The time-dependent perception of a RIR is a fundamental key to the idea of combining reproduction techniques. Therefore, a brief overview of the definitions of the time sections and their relevance for specific perceptual qualities is provided.

2.4.1 Direct sound

The direct sound is the sound traveling along the "line of sight" and, consequently, the first sound wave to arrive at the listener. The information provided by the DS is mainly important for source localization. The so-called "law of the first

wave front" or precedence effect [Cre77; Kut16] describes the perceptual effect that even though reflected energy arrives at the listener from multiple directions, source localization depends on the energy arriving with the first wave. Sound waves arriving in a very short time section (few milliseconds) after the DS are "summed up" by the auditory system and produce a single, so-called "phantom source" as an auditory event.

2.4.2 Early reflections

Sound waves that arrive up to around 50 to 80 ms after the DS are called ERs. Each reflection with a wall attenuates the signal due to the frequency-dependent surface characteristics. As ERs typically are only reflected a few times, the resulting signal is similar to the DS but attenuated. The perception of ERs depends on the incidence angle, the intensity level, and the delay after the DS [Kut16]. ERs can enhance the overall loudness as they provide energy but do not change the perceived source position. Figure 2.4 illustrates the principles and the findings by Barron [Bar09] for perceiving ERs, for using music as a stimulus and a 40° lateral reflection. The x-axis indicates the delay of the reflection after the DS, the y-axis the intensity level of the reflection. For very short delays, the reflections are used for source summation (phantom source effect) by the auditory system. For reflections with very high intensity, the perceived source position also shifts. For delayed reflections with high intensity, a disturbance occurs, mainly perceived as a distinct echo. For very low levels, the region labeled "threshold", reflections become inaudible. Reflections around 20 ms can lead to tone coloration that is described as "sharpening of the timbre; it imparts a shrill, slightly metallic character to sound" [Bar09]. "Spatial impression" contributes to a perception of source width and perception of room (i.e., a three-dimensional sound presentation) and relates to lateral energy. Besides these mentioned effects, ERs contributes to intelligibility of speech and clarity of music [Vor20].

2.4.3 Diffuse decay

The latter, reverberant part of reflections is mainly described by its exponential energy decay and is therefore called DD in this thesis. In general, the diffuse nature of these densely arriving reflections results in a stable perception that makes differences between different listener positions and orientations inaudible within boundaries, as the assumption for sufficient diffuseness might not stand for positions close to walls or obstacles or special rooms with directional reverberation (long corridors or highly absorptive surfaces at only one room direction). The late, reverberant part of the impulse response contributes to the overall perception

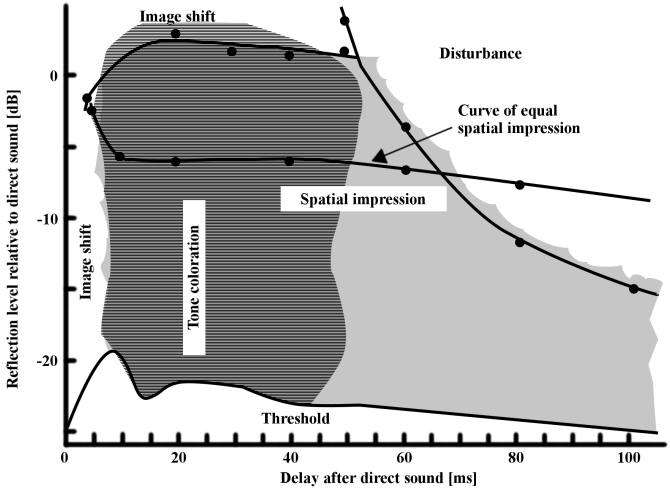


Figure 2.4: Changes in perception due to different energies and time of arrival of a single, lateral side reflection (40°), after [Bar09]. Music was used as a stimulus. A reflection that arrives in the first milliseconds after the DS or has very high energy can lead to a shift in perceived source location. A reflection with energy below the threshold line is inaudible, while a very late reflection with high energy is perceived as a disturbing echo. Tone coloration is audible due to the comb filter effect. The spatial impression is a result of the lateral incidence of the sound and can be perceived as widened source extension.

of reverberation time and especially listener envelopment [Kut16; Vor20], where listener envelopment again is linked to the lateral arriving energy.

2.4.4 Mixing times

In a RIR, the density of reflections increases over time [Kut16] and results in an increased physical and perceptional diffuseness. The time between the direct sound and the transition from ERs to the DD is called mixing time. Note that none of the terminologies are defined in hard requirements or limits, and the transition is rather smooth, with both concepts of ERs and DD overlapping in time (as indicated in Figure 1.1). The diffuseness of the sound field can physically be defined by arbitrary incidence of sound, as well as a position-independent sound pressure. Consequently, the perceptual definition of the sound field after the transition time can be formulated as a change in position or rotation of the user with no change in the perception of the sound field, and therefore, single reflections can neither be perceived in time or direction of arrival nor in intensity.

The energy decay as such, on the other hand, can still be perceived. Naturally, the physical and perceptual mixing times do not have to match. This thesis focuses on the perceptual definition. In classic room acoustics theory, which focuses on concert halls, the early energy is defined to end somewhere between 50 ms and 80 ms, depending on different factors, e.g., stimulus used (speech versus music). A comparative investigation on different attempts to estimate the mixing time was done by Lindau et al. [LKW12]. They compared listening test results of nine different virtual rooms with different models to predict the mixing time. As good estimator for the mixing time t_m in milliseconds, depending on room volume V and surface area S, is stated as:

$$\mathrm{t_m} = 20 \cdot \mathrm{V/S} + 12 \ \ [ms] \tag{2.1}$$

The study further concludes "that a time interval corresponding to about two mean free path lengths, i.e., on average two orders of reflection, and a reflection density of less than 200 s^{-1} is perceived as diffuse even by trained listeners."

2.5 Rendering and room simulation

Virtual scenes are defined by their input data. Rendering processes this information to return an output signal that can be processed for audio playback. In this thesis, the rendered audio outputs will be either a two-channel binaural audio signal or a multichannel signal. The binaural stream can be either used for headphone reproduction (ideally compensating the headphone transfer function (HpTF)) or has to be further processed with CTC filters. For Vector-Base Amplitude Panning (VBAP), the rendered multichannel signal is the direct loudspeaker signal. For Higher Order Ambisonics (HOA), this signal depends on the truncation order used and has to be decoded to the loudspeaker setup on the reproduction side. Hence, the output of the rendering process includes all information about the acoustic virtual scene and its sound field. The task of the reproduction is to deliver this information as unchanged as possible to the listener. Typical input data for a virtual acoustic scene is:

- source position, orientation, directivity
- listener position, orientation, directivity (e.g., HRTF)
- the geometrical description of the environment (e.g., the room, obstacles)
- acoustic properties of the geometric surfaces (e.g., absorption and scatter coefficient)
- definition of the medium, which is typically air (e.g., density and humidity)

To render the sound field, simulations have to be run to approximate the resulting sound field at the listener's position. Simulation methods like finite and boundary element method include wave-based effects, but are far from calculation times that are considered real-time. A computationally more efficient approach is using geometrical acoustics. In geometrical acoustics, sound propagation is approximated by particles (or rays) along the normal vector of the emitted waves. The software Room Acoustics for Virtual Environments (RAVEN) [SV11; Sch11] used in this thesis combines two approaches to simulate sound propagation in a room. The image source approach [AB79] calculates the sound path along specular reflective surfaces. The order of the image sources describes how often a path was reflected by a wall before it arrives at the listener. The method is deterministic and delivers distinct information about time, energy, and direction of arrival of single reflections. Nevertheless, the calculation time for image sources increases exponentially with the calculated order. Therefore, for higher-order reflections, Ray Tracing is used. Ray Tracing is a stochastic approach where a sufficient number of rays propagate from the source. The volume of the listener is expanded to allow for a suitable hit probability. Besides lower calculation times for longer impulse responses, this approach includes scattering from the surfaces. For both approaches, air absorption is calculated. For detailed information about geometrical acoustics simulation see [Vor20], and for the specific implementation of RAVEN see [SV11; Sch11]. Shortcomings of this approach are missing modal behavior for lower frequencies, especially for rooms with smaller volumes, as well as missing diffraction.

2.6 CTC - Crosstalk cancellation

A crosstalk cancellation (CTC) system is a binaural, loudspeaker-based method that aims to reproduce a binaural two-channel input signal at both ears separately and might be visualized by the term "virtual headphone". As spatial audio perception relies on inter-aural *differences*, the channel separation is crucial. Yet, when using a loudspeaker, e.g., the left ear, the signal will arrive at both ears. To compensate for the signal arriving at the right ear, a second loudspeaker can be placed to cancel the intended signal. Yet, this signal will again crosstalk to the left ear, which then has to be canceled out by the first loudspeaker. For this principle to be stable, the signal paths (i.e., left loudspeaker to left ear and right loudspeaker to right) have to have higher amplitudes than the crosstalk paths (i.e., left loudspeaker to right ear and right loudspeaker to left ear).

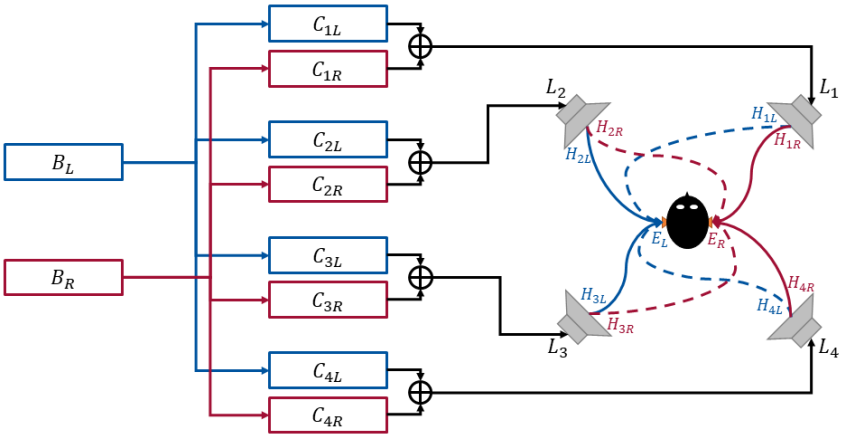


Figure 2.5: Schematic of a CTC system with N=4 loudspeakers. The input signal B_L and B_R are fed to the CTC filter network and summed up to loudspeaker signals **l**. The transfer paths from the loudspeaker to the ears are described by the matrix **H**. The signals then arrive at the two ears indicated by the ear signals **e**. Signal paths relating to the left binaural input or the left ear side are indicated in blue, for the right side in red color. Crosstalk paths are indicated by dashed lines.

Different approaches to solve this problem exist and are summarized in [Mas12]. This thesis focuses on real-time application in interactive environments and uses the closed solution in the frequency domain, utilizing finite impulse response (FIR) filters to achieve low computational costs and the use of more than two loudspeakers to result in stable filters for any user orientation.

Figure 2.5 illustrates the principle of the crosstalk cancellation system. The binaural input $\mathbf{b} = [B_L\ B_R]$ is fed to an electrical filter network **C**, then summed up to the N loudspeaker signals $\mathbf{l} = [L_1\ L_2\ ...\ L_N]$, with N=4 in the picture. The transfer functions from the loudspeaker to the two ears are described in **H**. The signals arriving at the ears are denoted by $\mathbf{e} = [E_L\ E_R]$. It should be noted that all elements are complex frequency vectors. The objective of the CTC system is to compensate the influence of the acoustics transfer paths **H** by using adequate filters **C** so that:

$$\mathbf{H} \cdot \mathbf{C} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{2.2}$$

The acoustic transfer paths for N loudspeaker is of size 2xN and consists of:

$$\mathbf{H} = \begin{bmatrix} H_{1L} & H_{2L} & ... & H_{NL} \\ H_{1R} & H_{2R} & ... & H_{NR} \end{bmatrix} \tag{2.3}$$

The first index indicates the loudspeaker, the second one the ear side of the listener. To solve the problem stated in equation 2.2, $\mathbf{H}$ has to be inverted so that:

$$\mathbf{C} = \mathbf{H}^{-1} \tag{2.4}$$

To invert $\mathbf{H}$ the Moore-Penrose Inverse is used. The inversion of $\mathbf{H}$ leads to non-causal filters. To avoid this problem, the CTC filters are delayed (i.e., time shifted) and account for the traveling time of the sound wave from loudspeaker to listener. The crosstalk cancellation matrix is then described by a Nx2 matrix, where the first index relates again to the loudspeaker number and the second to the binaural input:

$$\mathbf{C} = \begin{bmatrix} C_{1L} & C_{1R} \\ C_{2L} & C_{2R} \\ ... & ... \\ C_{NL} & C_{NR} \end{bmatrix} \tag{2.5}$$

The solution of the problem requires a division by $\mathbf{H}$, which leads to very high gains for frequencies close to zero (notches in an HRTF) in $\mathbf{H}$. These notches are usually located at higher frequencies, where smaller mismatches of the estimated transfer function compared to the real transfer function lead to audible artifacts and high-frequency ringing. Even under free-field conditions, mismatches occur due to the latency of the system and listener rotation during the calculation of the filters and playback, and the limited accuracy of HRTF measurements for higher frequencies [Rie98; Koh+21]. Besides these mismatches, the loudspeakers are limited in gain and cannot provide extreme energy levels. To avoid these effects, a regularization is typically applied on the HRTFs, indicated by μ. Additionally, for loudspeaker systems with more than two loudspeakers (N>2), multiple solutions exist and the system is under-determined. As an additional constraint for the solution, the minimization of the energy in the loudspeaker signals is achieved by using least-squares minimization. The solution is then given by [Mas12] as:

$$\mathbf{C} = \mathbf{H}^{*} \left(\mathbf{H}\mathbf{H}^{*} + \mu\mathbf{I}\right)^{-1} \tag{2.6}$$

where * denotes the complex conjugate matrix and $\mathbf{I}$ the identity matrix.

2.6.1 Perceptual aspects

Binaural sound reproduction over a CTC system is prone to sound coloration [Gua+07]. The mismatch of the calculated transfer paths from the loudspeakers to the ears can result in comb-filter-like coloration, especially for high frequencies where, due to notches in the HRTF, higher gains occur in the CTC filter. For CTC systems, where the transfer path is dynamically updated to adjust to the listener's changes in position and orientation, a free-field measured HRTF is used. In addition to neglecting the room acoustics around the loudspeakers, these measurements are subject to a certain variability [Rie98; Koh+21], and the calculated transfer path is never perfectly accurate. Furthermore, the tracking system to acquire the listener's position and orientation has a limited spatial and temporal resolution. As this thesis is aims to provide a low inhibition threshold for entering the system and avoiding the requirements of an anechoic chamber and cumbersome measurements, free-field artificial head HRTF measurements are used that affect the CTC system on two levels: the binaural rendering and the CTC filters. As shown in previous studies [MMF13; Mas12], non-individual CTC filters perform similarly to slightly mismatched but individualized CTC filters, with the exception that sources located in the rear hemisphere produce higher back-to-front confusions when simulating two loudspeakers in a headphone experiment. On the rendering side, the use of non-individualized HRTFs can lead to front-to-back confusions and increased errors in distance perception [Møl+96]. Further studies [BWA01] using speech stimuli show that confusion rates can significantly be reduced by adapting the sound presentation to the listener's movements, i.e., by dynamic reproduction. The perception of externalization, on the other hand, is mainly influenced by adding reverberation to the scene, and the effect size of using individual HRTFs is less important. A further overview can be found in literature [Bla13]. Findings on localization accuracy mainly exist for setups using two loudspeakers or loudspeakers simulated in a headphone experiment. In [Len06], median values of localization accuracy are close to zero, but results spread within a range of $\pm 25°$. Compared to VBAP and HOA, the CTC system has a reaction time towards listener movements, as it has to determine the new listener position, calculate the CTC filters, and update these filters (see also Section 6.1.2).

2.7 VBAP - Vector-Base Amplitude Panning

Vector-Base Amplitude Panning (VBAP) is the consequent extension of Stereo panning and is presented in [Pul97]. By adding a third speaker, the virtual sound sources can be panned within a triangle formed by three speakers. For full 3-D

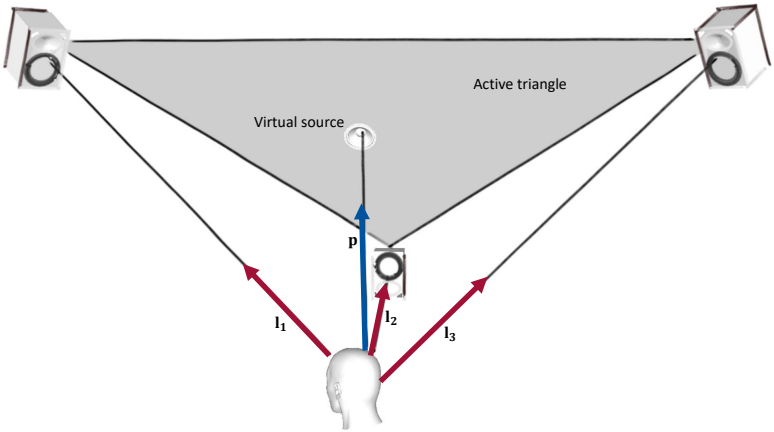


Figure 2.6: Concept of Vector Base Amplitude Panning. The vectors $\mathbf{l_1}$,$\mathbf{l_2}$ and $\mathbf{l_3}$ point toward the active speaker and span a vector base in which the virtual source position $\mathbf{p}$ is described. The picture is similar to [Pul97].

reproduction, a convex hull of loudspeaker triangles has to surround the listener. For the calculation of the speaker weights in an active triangle, the vectors from the center of the loudspeaker array towards the speaker, $\mathbf{L} = [\mathbf{l_1 l_2 l_3}]$, are used to span a vector-base (see Figure 2.6) and therefore have to be unit-length. The loudspeaker weights $\mathbf{g} = [g_1 g_2 g_3]$ can then be obtained by using the vector-base to describe the sound source position using the unit-length vector $\mathbf{p}$. Pulkki's Formula then forms into:

$$\mathbf{g} = \mathbf{p}^T \mathbf{L}_{123}^{-1} = [\ p_1\ p_2\ p_3\ ] \begin{bmatrix} l_{11}\ l_{12}\ l_{13} \\ l_{21}\ l_{22}\ l_{23} \\ l_{31}\ l_{32}\ l_{33} \end{bmatrix}^{-1} \tag{2.7}$$

The squared weights $\mathbf{g}$ are scaled to a constant C by

$$\mathbf{g}^{scaled} = \frac{\sqrt{C}\mathbf{g}}{\sqrt{g_1^2 + g_2^2 + g_3^2}} \tag{2.8}$$

VBAP results in three different numbers of active loudspeakers for a point source in a free-field environment:

- Virtual source in the direction of a real speaker: One loudspeaker is active.
- Virtual source on the edge of an active triangle: Two loudspeakers are active.
- Virtual source within an active triangle: Three loudspeaker are active.

Where the term active is defined as a weighting factor higher than zero. The VBAP approach is incapable of providing information about source distance apart from those sources positioned on the surface spanned by the loudspeaker triangles. Yet, appropriate rendering changes the intensity of the direct sound as well as direction, time, and intensity of the reflections for different source distances, enabling different distance perception.

2.7.1 Perceptual aspects

VBAP utilizes a minimum number of active loudspeakers for panning and is considered very robust [ZF19] against listener movement. Similarly, as for Stereo reproduction, the superposition of the sound emitted by the active loudspeaker at the ears of a listener is frequency-dependent. While the low frequency portion sums up coherent, high-frequency parts sum up energetically, resulting in an emphasized perception of low frequency for coherent input signals and dry reproduction environments. For mid and high-frequency, comb filter effects might occur due to slightly different traveling paths from each loudspeaker to the ears. This behavior further changes due to the sound source position in relation to the nearest loudspeaker in the active triangle and the number of loudspeakers being active. Coloration can therefore be expected to some degree between positions and for moving sources. Pulkki stated in his original paper [Pul97] that localization blur depends on the source location in the active triangle and the size of the triangle, which is related to the distance of the source to the closest loudspeaker. In his further investigations [Pul01] he showed that the perceived source location is individual and varies up to $\pm 15°$ in elevation for a wide span loudspeaker triangle as used in this thesis. For horizontal deviations, the spreading decreases to half. These effects, of course, are dependent on the loudspeaker arrangement and the source position relative to it.

2.8 HOA - Higher Order Ambisonics

Like any signal in the time domain can be approximated by a superposition of sinusoidal signals, the sound field direction can be approximated by a series of spherical harmonics (SH). Ambisonics decomposes the sound field using these SH and uses the superposition of SH components provided by the loudspeaker to match the decomposed sound field. A detailed overview of the technique and the specific mathematics is given in [ZF19]. The following observations rely on plane wave propagation, hence, both virtual source and loudspeaker have to be placed at a minimal distance where far field conditions apply. Furthermore, the virtual source is required to be outside the loudspeaker array.
The real valued SH coefficients can then be found using the SH function:

$$Y_n^m(\theta,\varphi) = N_n^{|m|} P_n^{|m|}(cos(\theta)) \begin{cases} sin(|m|\varphi), m < 0 \\ cos(|m|\varphi), m \geq 0 \end{cases} \tag{2.9}$$

With φ being the azimuth and θ being the elevation angle for the incoming sound wave, as introduced in 2.1. Consequently, the sound wave is traveling in the $(-\theta, -\varphi)$ direction. m describes the degree, n the order of the SH function. The associate Legendre functions are defined as:

$$P_n^{|m|}(cos(\theta)) = (-1)^m (1 - cos^2(\theta))^{m/2} \frac{d^m}{dx^m}(P_n^0(cos(\theta))) \tag{2.10}$$

N is a function to normalize the associated Legendre function and the full normalization (N3D) is:

$$N_n^m = (-1)^m \sqrt{\frac{(2-\delta_m)(2n+1)}{4\pi} \cdot \frac{(n-m)!}{(n+m)!}} \tag{2.11}$$

Where the δ_m being 1 for $m = 0$ and 0 for all other cases. Multiple normalization approaches exist, see [Car17] for an overview.

Ambisonics is divided into encoding the sound field that is captured (and that is to be reproduced) and decoding it onto a loudspeaker array. A virtual source, or more specific, its emitted plane wave can then be encoded using SH up to a truncation order N:

$$y_{vs} = \sum_{n=0}^{N} \sum_{m=-n}^{n} Y_n^m(\theta_{vs}, \varphi_{vs}) \tag{2.12}$$

This results in $(N+1)^2$ coefficients:

$$y_{vs} = [Y_0^0(\theta_{vs}, \varphi_{vs}), Y_1^{-1}(\theta_{vs}, \varphi_{vs}), Y_1^0(\theta_{vs}, \varphi_{vs}), ..., Y_N^N(\theta_{vs}, \varphi_{vs})]^T \tag{2.13}$$

The signal driving a source is then multiplied with these coefficients and results in an audio stream with $(N+1)^2$ channels. The truncation of the SH leads to unwanted side lobes which decrease the sweet spot area of the system. While the use of higher orders decreases the magnitude of the unwanted side lobes, it also increases the number of channels to be used and, more importantly, the minimum number of loudspeakers needed for a stable reproduction using mode matching decoding. As the number of loudspeakers present restricts the use of high truncation orders, a weighting of each order (window function over orders) can be applied to suppress the cut-off effect towards higher orders. In general, attenuating higher orders is a trade-off between narrowness of the main lobe pointing towards the virtual source and side lobe suppression. The most prominent approach is the maximization of the panning energy towards the source, the max-r_E decoding [Dan00; DRP98], which results a good balance in this trade-off. The weights needed for each order are indicated by a_n and for each order $2n+1$ degrees exist for which the same weight needs to be applied so that the weights can be applied using a diagonal matrix:

$$y_{vs,weighted} = y_{vs} diag\{a_0, a_1, a_1, a_1, a_2, ..., a_N\} \tag{2.14}$$

The encoding of a loudspeaker array with L loudspeakers is:

$$\mathbf{Y}_L = [y_{l1}, y_{l2}, ..., y_L] = \begin{bmatrix} Y_0^0(\theta_{l1}, \varphi_{l1}) & Y_0^0(\theta_{l2}, \varphi_{l2}) & ... & Y_0^0(\theta_L, \varphi_L) \\ Y_1^{-1}(\theta_{l1}, \varphi_{l1}) & Y_1^{-1}(\theta_{l2}, \varphi_{l2}) & ... & Y_1^{-1}(\theta_L, \varphi_L) \\ ... & ... & ... & ... \\ Y_N^N(\theta_{l1}, \varphi_{l1}) & Y_N^N(\theta_{l2}, \varphi_{l2}) & ... & Y_N^N(\theta_L, \varphi_L) \end{bmatrix} \tag{2.15}$$

The weighted superposition of the sound field encoded by the loudspeaker should then match the encoded sound field of the virtual scene by adequately weighting each loudspeaker with a real valued factor $\mathbf{g}$, including negative values (i.e., 180° phase shifting), so that:

$$\mathbf{Y}_L \mathbf{g} = diag\{\mathbf{a}_N\} y_{vs} \tag{2.16}$$

The weights for the loudspeaker can then be calculated by using the pseudo inverse denoted by $()^\dagger$:

$$\mathbf{g}_L = \mathbf{Y}_L^\dagger diag\{\mathbf{a}_N\} y_{vs} \tag{2.17}$$

2.8.1 Perceptual aspects

HOA and VBAP share the same behavior for coloration of input signals (see Section 2.7.1) as well as source distance information. Compared to VBAP, Ambisonics results in a more homogeneous, i.e., less source position dependent, reproduction especially for free field sources. Due to the limitations for the truncation order by the number of available loudspeakers, the main energy lobe pointing towards the source is wider and even further extended by the side lobe suppression. This leads to an emphasized source width perception and localization blur. Besides dependency on the loudspeaker arrangement, various decoding options for Ambisonics exists [ZF19] which makes comparisons more difficult. An overview of experiments on 2-D Ambisonics decoding can be found in [ZF19] and [Fra14]. For 3-D arrays, publications can be found using different decoding strategies and number of loudspeakers. In [TAK17] a first order system with six loudspeakers is used and a mean localization error of 25.6° found with a range above 35° for both, 50 percent and 95 percent confidence interval. For third order Ambisonics with 26 loudspeakers, a mean error of 15.44° was found and a range above 25°. For a fifth order system, Gerken et al. [GHG24] found increased localization accuracy with mean values below 5° in the horizontal and up to 30° in the median plane for elevated sources. The interquartile range is below 20°.

3

Setup and implementation

During this thesis, the Virtual Reality Laboratory (VR-Lab) was used to conduct listening tests. This chapter describes the room and the loudspeaker array present, as well as common procedures and assets used in the different listening tests. Additionally, the actual implementation for each reproduction technique is documented. For signal processing, measurements and analysis, the ITA-Toolbox [Ber+17] for MATLAB was used.

3.1 Virtual Reality Laboratory

The VR-Lab is located at the Institute for Hearing Technology and Acoustics, RWTH Aachen University, Germany (IHTA) and provides the technical environment for testing, conducting listening experiments as well as experiencing (acoustic) virtual reality via an head-mounted display (HMD) or a tracked, stereoscopic screen display. The room and setup is depicted in Figure 3.1. The loudspeaker array is discussed in Section 3.2 below. For a suitable, room-acoustical environment, the ceiling is acoustically treated, and the walls are covered with curtains during sound presentation. For the room acoustical characteristics below 200 Hz, four plate and Helmholtz resonators are placed in the corners of the room. In the upper part of the picture, a rail system can be seen which allows separating the control part of the room from the actual presentation part with curtains. Additionally, the textile projection screen can be closed off with curtains. After closing the curtains, the measured reverberation time is then about $T_{30,mid} = 0.15s$. Detailed data about the room acoustical optimization, room acoustical measurements and ambient noise levels can be found in [Pau22]. The Box in the ceiling houses the projectors and provides two two-channel audio outputs as well as two one-channel audio inputs.

The tracking system consists of six optical cameras (Flex13, Optitrack, Naturalpoint, Inc., Oregon, USA) in the ceiling (red colored in Figure 3.1) that can track users orientation and position with an update rate of 120 Hz and an accuracy of less than 1 mm depending on the calibration of the system. Figure 3.2 shows the white, reflective tracking balls which are equipped to a frame of glasses without lenses. The tracking point is corrected to the center of the inter-aural axis, as explained in Section 3.7.1 below. Additional to the projection system, a VIVE (HTC Corporation, Taoyuan, Taiwan) HMD system with two tracking cameras is installed. In this thesis, the main PC is a consumer desktop PC with an Intel Core i7-4790, 3.60 GHz and 16 GB RAM using a Windows 10 Enterprise operating system and an NVIDIA Geforce GTX 285 GPU with 1 GB GDDR3 memory. The second PC used for tracking has the same setup except for the GPU, which

Figure 3.1: Picture of the VR-Lab and equipment setup. Twelve loudspeakers in three horizontal rings with different elevation can be seen. Visible acoustic measures are the acoustic ceiling as well as the curtains, which can also cover the screen. The box in the ceiling contains two projectors for visual stereoscopic presentation via polarization. The red cameras in the ceiling are part of the tracking system. The user is equipped with tracked glasses and a (tracked) game pad for interaction. The forward direction is towards the screen. *Copyright at Institute for Hearing Technology and Acoustics, RWTH Aachen University, Germany*

is an NVIDIA Geforce GTX 580 with 1.5 GB GDDR5 memory. As input devices a tablet (Surface 3, Microsoft, Washington, USA) was used via remote desktop providing a listening test GUI. Alternatively, a game pad was used to pan virtual sources, as shown in picture 3.2.

3.2 Loudspeaker array

The VR-Lab provides twelve loudspeakers as spherical segments on three different rings, as depicted in Figure 3.1. The horizontal ring consists of four three-way loudspeakers (O300, Klein+Hummel, Ostfildern/Kemnat, Germany) arranged in 90° spacing starting at an angle of $\varphi = 45°$ from the frontal direction which is pointing towards the screen. The upper and lower ring are elevated at $\theta = \pm 30°$. In each ring four two-way loudspeakers (O100/O110, Klein+Hummel, Ostfildern/Kemnat, Germany) are spaced 90° horizontally with a starting angle of $\varphi = 0°$. All loudspeakers are placed at a distance of 2.3 meters from the

center point of the array and orientated towards the center. Additionally, a subwoofer (O800, Klein+Hummel, Ostfildern/Kemnat, Germany) is available. The arrangement of loudspeakers and their coordinates can also be seen in Figure 3.3.

Figure 3.2: Picture of the tracking glasses used in the VR-Lab, the polarization lenses are removed. The tracked point is the geometric center of the triangle spanned by the four white tracking markers. This point was shifted in post-processing to match the center point of the aural axis. The game pad was used as an input device to pan virtual sources.

3.3 Artificial head and HRTF

The artificial head used in this thesis was developed at the IHTA [Sch95] and is used for acquiring the HRTF datasets for binaural synthesis and CTC filter generation. It uses two MK 2H microphones (Schoeps GmbH, Karlsruhe, Germany) and also consists of a torso part. It was further used to acquire sound pressure levels at both ear drums during sound reproduction and as validation method during the CTC reproduction. The HRTF dataset used in this thesis was measured with a measurement arm equipped with a loudspeaker, which measured the head with a resolution of 1° both in azimuth and elevation at a distance of 1.86 meters. The measurements were done in a hemi-anechoic chamber with time windowing set to avoid any influence of the ground reflection. Each direction was saved with 256 samples at a sampling rate of 44100 Hz. See [Ric19] for more details on the measuring device and procedure.

3.4 Crosstalk cancellation

The CTC system is realized using the four loudspeaker in the horizontal plane. The filters are calculated with a regularization of 0.001 to avoid sharp notches. A free-field measured HRTF dataset was used to estimate the transfer function between the loudspeaker and the listener's ear, as described in Section 3.3. Time shifting of the filters is applied, and the overall filter length needed to be estimated by the following parameters: The distance of the loudspeaker (2.3 meters) at a speed of sound of 343 m/s can be calculated to a delay of 295.7 samples. Allowing a stable reproduction with a radius of 2 meters, the filters need a reserve of about 256 samples in both directions, i.e., 512 samples. 256 samples of the head-related impulse response (HRIR) itself have to be added, and some buffer for windowing the filter to avoid hard onsets. All in all, a 2048 sample filter was chosen. A time shift of half the filter length plus 300 samples for the delay was selected to avoid non-causalities which result in filter peaks in the middle of the filter when the listener is positioned in the center of the array, i.e., 1024 samples. The used HRTFs were adjusted in level and delayed according to the listeners' position. Attempts to include first order reflections into the approximation of the transfer function for the CTC filter calculation showed decreased localization performance and increased coloration [Koh+16; KSV17] and was therefore discarded.

3.5 Vector-Base Amplitude Panning

VBAP is implemented as suggested in [Pul97] with a scaling constant of $C = 1$. Loudness adaptation due to changes in distances is done in the rendering. The convex hull is triangulated using a Delaunay triangulation (as provided by MATLAB), thus, maximizing the smallest angle in the resulting triangles. The resulting triangles can be seen in Figure 3.3. Additional to the source distance, the distance of 2.3 m from the loudspeaker to the center of the loudspeaker array is compensated frequency independent with a fixed constant for each loudspeaker:

$$g_{dist} = 20 \cdot log_{10}\left(\frac{2.3m}{1m}\right) = 7.23dB \tag{3.1}$$

As the listener is assumed to be seated in the center of the array, no compensation for displacement of the listener from the array's center is implemented.

3.6 Higher Order Ambisonics

Ambisonics encoding is realized according to Section 2.8 with full normalization (N3D). Decoding is done using maximization of the energy vector r_E [Dan00] and a mode matching approach using the Moore-Penrose pseudo-inverse. As the upper and lower cap of the sphere are lacking loudspeakers, two virtual loudspeaker positions were added at direct upwards and downwards position ($\vartheta = \pm 90°$) for the decoding. They were evenly routed to the upper and lower ring to ensure mathematical stability and avoid loudness changes in these directions [ZF19]. To route the virtual loudspeaker to the four existing ones, the weights were calculated so that the root of the squared sum equals one, i.e., with a factor of 0.5. Again, changes in distance were realized by the rendering and no compensation for a displaced listener is implemented. The distance of the loudspeaker was compensated in the same way as for VBAP, see equation 3.1.

3.7 Listening tests

The conducted listening tests share some common procedures and set-ups. To avoid redundant description, they are explained in the following section.

3.7.1 Positioning of participants

The reproduction techniques used rely on a properly positioned participant. While VBAP and HOA reproduction rely on a listener centered in the midpoint of the array, the CTC requires accurate information about the listener position and orientation. The so-called "rigid body" is the geometrical mid-point of all tracking balls and has to be adjusted to match the interaural axis of the participant. As the fitting of the glasses is individual, the adjustment had to be done for each participant. As VBAP and HOA, in their specific implementations, are more robust against listener movement and the CTC system relies on accurate tracking data, a perfect pin down of the participants is not needed and discarded in favor of a practical solution where listeners are free to move their heads in some boundaries. To match the center of the interaural axis of the participants with the center of the array a three axis laser was positioned so that the horizontal plane matched the center of all four horizontal loudspeaker and the vertical axis was aligned with the loudspeaker elevated at the direct left and direct right side (azimuth: $\varphi = \pm\ 90°$, elevation: $\theta = \pm\ 30°$). The participants were positioned so that the crosspoint of these lines were aligned with the center of the right ear canal entrance (the head of the participants shadowed any line on the left ear). This step calibrates the height positioning (y-axis) and the front-back translation

(z-axis). During this procedure, the participants were instructed to look at a visual white cross on a black background that was shown in the direct frontal direction to control head orientation (azimuth and elevation angle). An additional vertical line was projected by a second laser matching the elevated loudspeaker in frontal and rear direction (azimuth: $\varphi = \pm\ 0°$, elevation: $\theta = \pm\ 30°$). This vertical line together with the horizontal line crossed at the position of the visual cross. The vertical line was used to calibrate the left right translation (x-axis). In practice, the x-axis calibration did not change between participants, so the actual calibration task narrowed down to match ear canal entrance to the laser's cross point. For the listening tests and the positioning procedure, a non-rotatable chair was needed that provides enough height to lift smaller people to an ear height of 1.34 meters. A food rest is needed to provide support for feet and legs during the long testing times. To further support the back and avoid unnecessary head movement, an adjustable back rest is needed. During the evolution of listening tests in this thesis, the final solution for positioning participants is a Kolberg 3105 chair (Kolberg Percussion GmbH, 73066 Uhingen, Germany) that is designed for harpists and fulfills all the requirements. To calibrate the tracking system, participants were instructed to look at the white cross in direct frontal direction. The tracking bodies orientation was then calibrated to zero (i.e., azimuth and elevation angle were defined as zero). Using a protractor, the horizontal laser line was used to achieve a 90° angle and the height of the tracking body was measured, followed by measurement along the horizontal measurement of the front back displacement along the z-axis. The tracked body was then corrected by this displacement in the tracking software. If, for some reason, the glasses were noticeably displaced during the listening test, the procedure can be repeated during the listening test, as the tracking software runs in parallel on a computer separate from the actual test.

3.7.2 Sound source positions

The choice of the tested sound source positions takes multiple aspects into consideration. While the listening tests were designed to investigate the specific parameter as general as possible, the duration of the tests is a limiting factor. Therefore, four different sound source positions were chosen. The right hemisphere was used due to a potential, more accurate perception in comparison to the left ear [Emm+88]. The elevation is limited by the loudspeaker array as sources above the upper ring, or below the lower one, lead to wide spanned VBAP triangles. Considering the active VBAP triangles and the distance to the nearest loudspeaker, different cases should be covered: a virtual source position in the direction of a real loudspeaker, on the edge of an active triangle (i.e., on a line

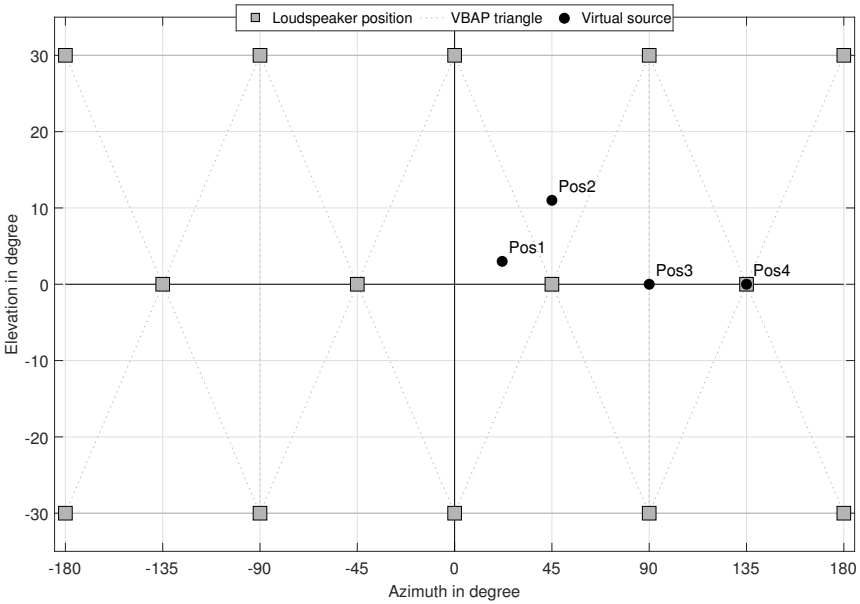


Figure 3.3: The Figure shows the loudspeaker array in the VR-Lab as elevation over azimuth. Loudspeaker are drawn as squares, dotted lines show the VBAP triangulation and solid black dots the virtual source positions (one to four from left to right) as indicated in the legend. Note, that the loudspeaker on the very left and very right, i.e., at $\pm 180°$, are the same.

between two loudspeaker) and within an active triangle. While the horizontal plane is typically more important for everyday scenarios, elevated sources emphasize the monaural cues more and are to be tested too. An attempt to fit all these different needs is made and resulted in the source positions shown in Table 3.1. Figure 3.3 shows the source positions in relation to the loudspeaker

Position	Azimuth [°]	Elevation [°]
1	-22	3
2	-45	11
3	-90	0
4	-135	0

Table 3.1: Positions of the virtual sources in the listening test.

array. Position 1 and 2 lie within an active triangle, with position 2 being the elevated source. Position 3 lies on an edge of an active triangle with high ITD

and ILD values. For position 4, the virtual source is placed in the direction of a loudspeaker used by the reproduction array. In all scenarios, these virtual sources are represented as point sources.

3.7.3 Virtual rooms

For the SAQI listening tests (Section 4.3 and 6.2.1) as well as for the subjective latency evaluation (Section 6.1.2) virtual rooms with increasing reverberation times were used. These virtual rooms were taken from the BRAS database [Bri+21; Asp+20] and are documented in detail, including photos of the real rooms. The rooms are named "CR2", "CR3" and "CR4" and will be defined as "small", "medium" and "large" room in this thesis. The small room is an abandoned seminar room which is empty and has a long reverberation time of about 1.5 seconds. To reduce this reverberation time, an absorbing surface was added to the floor of about 21.5 m^2 and the reverberation time adapted according to 5.2 using a target reverberation time of 0.5 seconds. The room volume is about 145 m^3 and the total surface area about 200 m^2. The reverberation radius calculates to 1.25 meters and the mixing time to 26.4 ms using equation 2.1. The medium room is the small hall of the Konzerthaus Berlin with a reverberation time around 1.5 seconds for mid-frequencies. The room volume is 2350 m^3 and the surface area about 1950 m^2. The reverberation radius calculates to 2.81 meters and the mixing time estimates to 35.6 ms. The large room is the Auditorium Maximum at TU Berlin, with a mid-frequency reverberation time above 2 seconds. The room volume is 8650 m^3 and the surface area around 5850 m^2. The reverberation radius is 3.55 meters and the calculated mixing time 41.6 ms. The frequency dependent reverberation times of the virtual rooms used can be seen in Figure 3.4. Visualizations of the room can be seen in Figures 4.5 and 6.6.

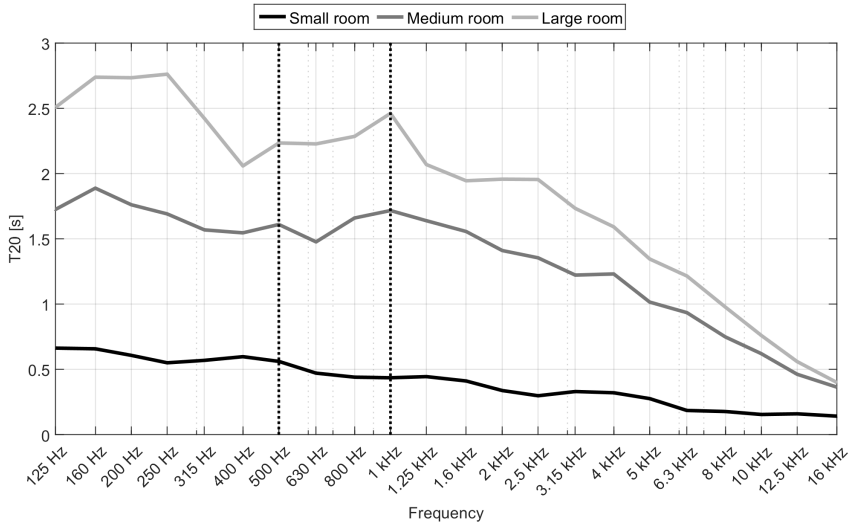


Figure 3.4: Reverberation times of the virtual rooms used in the listening test. The mid-frequencies typically used to estimate a mid-frequency, one value reverberation time, are indicated in dashed lines.

4

Perceptual qualities of reproduction systems

The combination of reproduction techniques is based on different strengths and weaknesses of the single reproduction methods. Findings on the performance of different reproduction methods rely on the specific implementation used, the testing method and its vocabulary, and, of course, the loudspeaker array (especially the number of loudspeaker) used and the room they are placed in. Last, the stimuli used have to cover the range of interest, in this case: different reverberation times of the virtual room. Consequently, the reproduction systems are tested towards their strength and weaknesses in terms of localization accuracy and audio quality (SAQI). The results will be used to determine which system should be combined. To ensure an inaudible transition between the reproduction techniques, a listening test regarding loudness differences between the system was done. The defined SAQI method and virtual rooms from the BRAS database [Bri+21; Asp+20] were used to make the findings more comparable for future studies.

4.1 Loudness

To ensure an inaudible transition between the reproduction techniques, they have to be adjusted in their perceived loudness to match each other. This section covers a subjective evaluation of the perceived loudness evoked by a virtual source presented over a reproduction system compared to those evoked by a reference loudspeaker at the same position. Multiple factors influence the perceived loudness of a reproduced source. The room in which the test is conducted in, and the loudspeaker array is set up affects the perceived loudness by adding its, position dependent, room acoustics. For VBAP and HOA a low frequency boost is expected due to the more coherent superposition of the loudspeaker signals at the listeners' ears compared to higher frequencies where sound superposes more energetically. Moreover, the relative position of the virtual source to the (nearest) loudspeaker is relevant, especially for VBAP where the number of active loudspeaker varies due to the position of the virtual source, which then influences the amount of low frequency boost. An important factor for the CTC system, on the other hand, is the HRTF used in the binaural synthesis, which has more fluctuations in higher frequency for different source positions (relative to the listener). The aim of the listening test is to develop a generally applicable, real-valued gain to minimize the loudness differences between the three reproduction techniques over all (tested) positions. A position or frequency dependent adaptation would alter the reproduction technique itself and is not in the scope of this investigation. Main parts of this chapter have also been published in [KV19a].

4.1.1 Method

The listening test aims at matching the loudness of the reproduction method to that of a reference loudspeaker at the position of the virtual source for an overall of four different positions. To obtain suitable amplifications, a 2-AFC procedure using the QUEST [Kin+94; WP83] approach was implemented that compares the reference loudspeaker against the virtual source reproduced at that position. The question displayed in the GUI was "Which stimulus is louder?" and possible answers: "first" and "second" (with randomized order of reproduction technique and reference source) given. The gain of the reproduction technique was then changed according to the QUEST calculation.

4.1.2 Setup

Four different positions for the virtual sources were chosen as described in Section 3.7.2 to cover the different influencing parameters mentioned above. Position three was slightly different and positioned at ϕ =-81° and ϑ =3° due to limitations in setting-up the reference loudspeaker. A 300 ms pulsed pink noise with 200 ms pauses was used to cover the full bandwidth of the reproduced sources and was repeated three times for an overall stimulus duration of 1.3 seconds. HOA and VBAP were calibrated to the reference by matching the average root mean square (RMS) sound pressure level (SPL) over all four positions while for the CTC system artificial head measurements were conducted and the mean of left and right ear were used. To achieve consistent loudness during a stimulus, the CTC filters were only calculated at the beginning of each relatively short stimulus presentation using the tracking glasses (3.7.1).

To minimize the visual influence of the room and the visibility of the loudspeaker array, the lights were turned off. Participants had to focus on a white cross in the frontal direction, which was highlighted by a lamp (see Figure 4.1). For tracking, participants were equipped with empty glasses that had tracking balls attached (see Section 3.7.1). The tracking point was corrected to the center of the interaural axis, i.e., the center point between the two ears. The GUI was provided via a tablet which controlled a Desktop PC on which the listening test was running in MATLAB, which then controlled Virtual Acoustics (VA). To adapt the participant's ear height to the center of the loudspeaker array, a platform was used on which a chair was placed. The backrest was used to achieve a more consistent position and orientation of the participants during the listening test. The set-up can be seen in Figure 4.1.

The task in the GUI stated: "Rate which stimulus you perceive as louder. Look at the white cross during playback and keep your head steady." To access the threshold in less steps, the QUEST [Kin+94] procedure was used with a fix number of 15 trials for each comparison. To avoid changes in the threshold during trials, the CTC filters were calculated for the participants' head position at the beginning of the trial. An informal interview was conducted after the listening test, asking for difficulties fulfilling or understanding the task or other annotations.

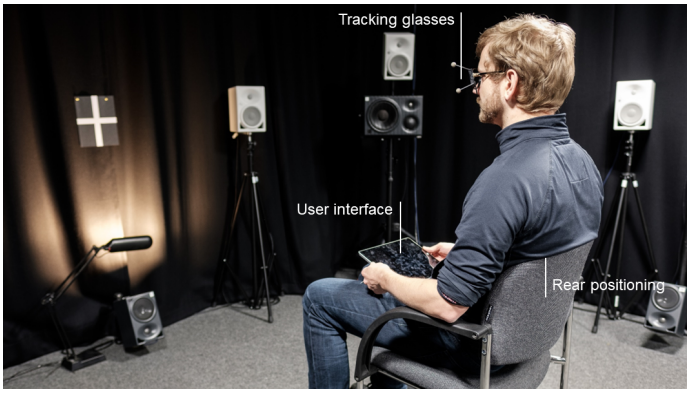


Figure 4.1: Setup of the listening test with a tablet for the GUI, empty glasses for tracking only and a chair for positioning the center of the interaural axis. The white loudspeaker were used for playback of the reference, the black ones are part of the reproduction systems. The lights of the room were turned off so that only the illuminated white cross was visible.

4.1.3 Results

20 normal-hearing participants with a mean age of 28.6 years ($\sigma = 4.35$) were tested for each position and reproduction technique in randomized order. Figure 4.2 shows the results of the listening test. The y-axis indicates the amplifications needed for the reproduction systems to excite the same perceived loudness for the participant than the reference stimulus did. For each position, the three reproduction systems are color coded as shown in the legend. The marker indicates mean values and the whiskers standard deviation. An ANOVA was used to statistically analyze the differences in the results (see Appendix A for detailed statistics). The interaction of position and reproduction technique was significant, and the simple mean effects were analyzed. Stars indicate the significance level. When no bracket is shown, it indicates a significant difference towards all other positions using the same reproduction technique (upper part)

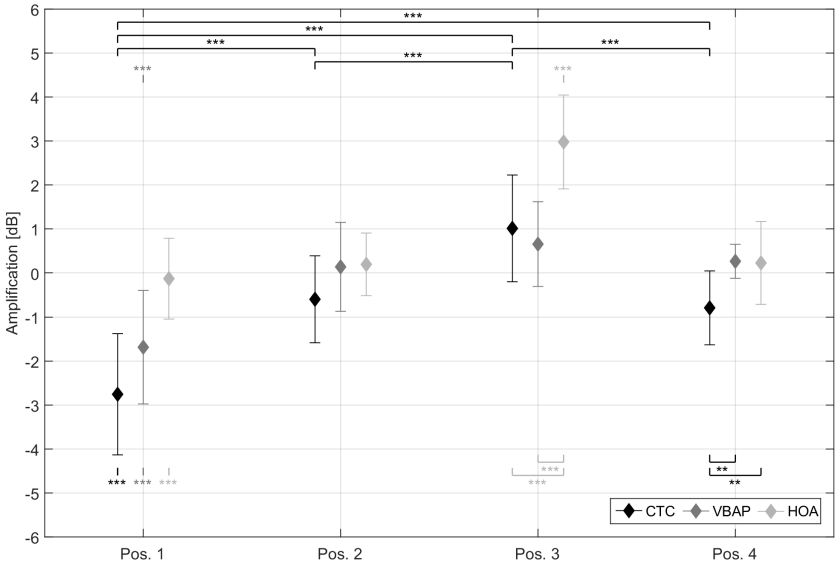


Figure 4.2: Results of the listening test as mean value and standard deviation. The amplification denotes how much a reproduction technique has to be amplified to achieve the same perceived loudness as the reference source. The differences found are in-between reproduction techniques and in-between source position. Significant differences between positions for one specific reproduction technique are indicated in the upper part of the plot. Significant differences between reproduction techniques in one specific position are indicated in the bottom part of the plot.

or towards all other reproduction techniques for the same position (bottom part).

For the effects within the reproduction techniques, positions one and three differ significantly from all other positions for CTC reproduction, while position two and four do not show significant difference between each other. For VBAP only the first and for HOA only the third position differs significantly from all others. For the effects within each position, position one shows significant differences between all reproduction techniques, whereas position three shows a significantly different behavior of the HOA system and position four for the CTC system. The range of the mean values can be found in Table 4.1 and is mainly influenced by position one (VBAP), three (HOA) or one and three CTC. Without these significantly different positions, the range of means is within about ± 1 dB. The

mean values range from −2.8 dB (Pos. 1, CTC) to +3 dB (Pos. 3, HOA) and result in a total range of 5.8 dB or ±2.9 dB. The maximum range in one reproduction technique was ±1.9 dB for the CTC reproduction. Furthermore, for all reproduction techniques, the results for position two and four are very consistent.

Method	min mean [dB]	max mean [dB]	mean range [dB]
CTC	-2.8	+1.0	3.8
VBAP	-1.7	+0.7	2.4
HOA	-0.1	+3.0	3.1

Table 4.1: Range of mean values per reproduction method.

During the post interview, participants stated a difference in perceived sound source location between reference source and virtual source (nine times, 45%). Moreover, coloration differences between reference and reproduced source made it difficult to compare the stimuli (six times, 30%).

4.1.4 Discussion

The listening test results show a general deviation of about 1 dB around each mean value for VBAP and HOA in the 50th percentile, which corresponds to just noticeable differences of 1 dB. The values for CTC range up to ±1.4 dB, around the mean for position one. Furthermore, the mean values as such are in the range of about ±1 around 0 dB, with one exception for each reproduction method: position one for CTC and VBAP and position three for HOA. Position two and four result similar findings even though position four is in the direction of a loudspeaker used by the reproduction methods and therefore both positions have different distances to the nearest loudspeaker. Both positions are almost on a cone-of-confusion and should result in similar ILD and ITD values. The lack of ILD and ITD and possibly a decreased source localization accuracy might also be a reason for the higher deviations seen for position one when played back over the CTC system. The decreased perception of loudness for position three during the HOA playback is unexpected, but might relate to the fact, that this is the position with the highest distance towards the nearest loudspeaker. For position four, the virtual source position is a position of a loudspeaker used by the reproduction systems. Therefore, during VBAP playback the signals are equal, resulting in a mean value close to 0 dB and very low spreading of the results, as can be expected. It should be noted that a general directional loudness dependency is avoided by the listening test design, as the reference is positioned in the same direction. Yet, the active loudspeaker themselves have different directions than the virtual and the reference source. The post interview revealed differences in perceived

source location between reference and virtual source, which is the focus in the subjective evaluation shown in Section 4.2. Changes in coloration are further investigated in Section 4.3, yet, stronger coloration effects are expected during the CTC playback and make it difficult for participants to compare loudness. Anyhow, it should be noted that the CTC filters do not change depending on the virtual source position in this test, however, the filters for the binaural synthesis do.

4.1.5 Conclusion

The range of deviation is reasonable when compared to a JND of 1 dB, and the CTC system is less homogeneous than the other two. The CTC system has an individual mismatch from each participant to the artificial head HRTF which increases the variability in loudness. For all systems, significant differences can be found in one position (two for CTC), but the limited amount of data points does not provide a conclusion whether these are outliers or systematics differences. ITD and ILD or source localization accuracy might be an influencing factor for the perceived loudness, especially for the CTC system. The position dependent variations of perceived loudness within each reproduction method, as well as the standard deviation, suggest that the adjustment of loudness does not need to be perfect. The final correction values are discussed in Section 6.1.1. Last, feedback from the participants revealed that coloration and differences in localization complicate the loudness comparison, which might be a reason for the increased spread CTC results.

4.2 Localization

Sound sources can be localized by the human auditory system up to a certain accuracy. Reproducing sound sources over a loudspeaker system can decrease this accuracy. In general, source localization accuracy mainly depends on the source position relative to the listener and the stimulus used. Besides the fact that each reproduction method provides different localization accuracy, they have different peculiarities that influence the provided accuracy. For VBAP the loudspeaker density is important as well as for HOA. Particularly for VBAP, it is imperative to consider the angular distance between the virtual source and the nearest loudspeaker of the reproduction array. Furthermore, the accuracy provided during HOA reproduction is also influenced by the truncation order and the decoding strategy employed, which is a compromise between robustness and localization accuracy (see Section 2.8) All three systems color the perceived sound to some degree and in different ways, which influences the localization

accuracy. The CTC system relies on a binaural synthesis that uses a given HRTF dataset. Furthermore, the CTC filters rely on an approximation for the transfer functions between loudspeaker and user with a limited accuracy. The filters are based on data provided by a tracking system with a limited update rate and accuracy, which typically does not incorporate the acoustic properties of the room. To gain knowledge about the localization accuracies provided by the systems in the specific loudspeaker set-up, the surrounding room and in the specific implementations (see Chapter 3), a listening test was conducted that compares the perceived localization accuracy against a (real) reference source. The findings will be used to select an appropriate reproduction system for reproducing the DS over the hybrid system. Findings of this chapter are also published in [KV21].

4.2.1 Method

To access the perceived sound localization, participants were asked to pan a virtual source, provided by a reproduction system, to the direction of a real reference source. The reference source was an additional loudspeaker positioned in the room, and participants could switch between the reference and virtual source as often as desired. The comparison against a reference source was chosen to cancel out general localization deviation and blur as shown in Section 2.2 and inaccuracies due to a pointing method. The GUI indicated whether the reference source was playing or the virtual source. Before the listening test, a training phase had to be completed. All loudspeakers in the room were labeled with letters. A signal was played back over one of the loudspeakers, and participants had to indicate which loudspeaker they perceived as "active". Last, the listening was ended with an informal oral interview regarding specific strategies used, artifacts heard and problems in understanding of and executing the task.

4.2.2 Setup

To match the inter-aural axis with the center of the loudspeaker array, wooden plates were put on a chair to adjust the participant's height. Tracking glasses were used, and the tracking system calibrated for each participant (see Section 3.7.1). Playback of the stimulus was stopped if the participants' orientation deviated more than 10° in the horizontal plane. This value is a balance between the benefits of a dynamic reproduction (including reduction of front-back confusions), listening to the stimulus without too many interruptions and keeping the intended source directions relative to the listener. The current orientation was indicated on the monitor, and a color indicator notified the participant when their orientation was outside limits. The source positions used can be found in Section 3.7.2 with

position four being a loudspeaker shared by the reproduction systems and the reference positions. Each position was tested three times and the stimulus was a pulsed pink noise with 300 ms signal and 200 ms pause with three repetitions. A monitor was mounted in front of the participants and interaction with the graphical user interface (GUI) as well as the virtual source panning was done via a game pad. The minimal step width was set to 1° which correlates to the resolution of the HRTF dataset used for the CTC system and binaural synthesis, see Section 3.4. To avoid front-back confusion and allow a more easy localization of the virtual source, the initial position of each trial was set to the front. Participants could reset the virtual source position if they lost track of the provided position during presentation over the reproduction system. To avoid confusion in the upper or lower hemisphere, where no loudspeaker for sound reproduction is present, the panning was limited to ±35° in elevation. For each position, the RMS sound pressure levels of an omnidirectional microphone in the center of the array were calibrated to the reference source for VBAP and HOA. For the CTC reproduction, the mean values of both microphones in an artificial head were used at a fixed frontal orientation.

4.2.3 Results

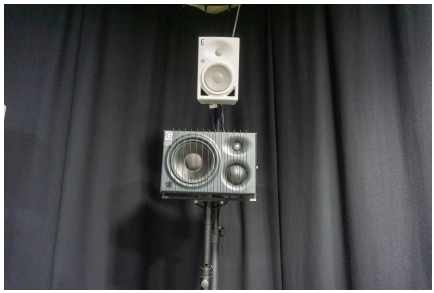


Figure 4.3: Reference position two and the front right loudspeaker (horizontal plane) of the loudspeaker array. Both position share the same azimuth angle, but differ eleven degrees in elevation.

The listening test was conducted with 29 normal-hearing participants, which had a mean age of 26.4 years ($\sigma = 4.59$). The duration of the test was about 40 minutes. In the training phase, the reference positions were easily identified by the participants except for reference source position two, which was placed eleven degrees in elevation above the loudspeaker two labeled 'G' of the array (see Figure 4.3). For this position, six participants (20,7%) indicated that the

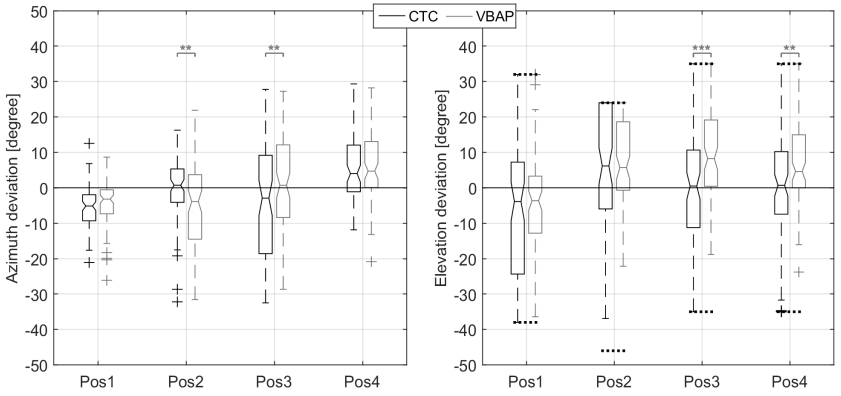


Figure 4.4: Results of the localization test. The left plot shows the deviations in the horizontal plane, the right one in the median plane. Dotted lines in the right picture show the limits of the panning.

presented sound was located at the black loudspeaker 'G', even though it was played back over the white reference loudspeaker 'E'.

During evaluation, the results for the HOA deviation in elevation were unexpectedly high, and an investigation revealed a misconfiguration which invalidated the HOA results. Especially for position two, where participants panned the virtual source to the upper limit and still stated the position could not be reached. A technical routing problem led to energy distribution from the virtual north pole loudspeaker to the reference loudspeaker at position three. Consequently, the more the source was panned upwards, the more energy was radiated from the horizontal plane. Even though the energy from the north pole loudspeaker is relatively low at the reference positions (26 dB below loudspeaker with maximum energy for position two, 36 dB and below for all other reference positions) the panning path of the participants might have a bigger influence. In conclusion, the results from the HOA reproduction were taken out completely. The implication of the hybrid system are discussed in Chapter 6.

Figure 4.4 shows the results of the listening test as angular deviation separated into azimuth and elevation. The box plot indicates the median value as a horizontal line, the notches indicate the uncertainty of the median, the boxes the range from 25th to 75th percentile. The whiskers indicate the full range, excluding outliers (indicated by '+') which are more than 1.5 times the interquartile range

below the 25th or above the 75th percentile. For each position on the x-axis, the two different reproduction methods are shown color coded according to the legend at the top of the figure. The horizontally dashed lines in the right plot indicate the limitations of the panning in elevation, which are calculated by subtracting the elevation angle of the source from the panning limits of ±35°. For example, for a reference source positioned at an elevation of 11° the maximum deviation upwards is 24° and downwards 46°. A two-way ANOVA was performed on the data (see Appendix A for detailed statistics) which showed interaction between position and reproduction method. Therefore, the simple means were investigated and are indicated by stars in the plot according to their significance level.

For the azimuth deviations, the spreading of the results around the mean value increases with more sidewards positions up to position three and decreases again for the more backward position. The deviations for the CTC system vary around 0°, with position one being offset to negative deviations and position four towards positive deviations. For the VBAP system, similar tendencies for position one and four can be observed. For the deviation in elevation, the spread of the results is, in general, higher and mostly ranges over the entire possible range. The CTC system is again distributed around 0°, with results spreading over the whole range and median values close to 0° for position three and four. The VBAP system shows similar values for position one and two, with less spreading and an offset towards higher elevations for position three and four. In the interview after the test, some participants stated that using only the right side was exhausting.

4.2.4 Discussion

The results from the training phase can be well explained by findings in literature (see Section 2.2, [Bla96])) where for a source in the median plane elevated by 36° the perceived sound source was stated at around 30° with a localization blur of about ± 10°.

Even though the listening test was designed to suppress effects that occur due to natural human source localization, similar effects can be observed. Using the interquartile range, the spreading is in the magnitude of those presented by Blauert [Bla96]. For the horizontal deviation, that is ±3.6° for the frontal and ±10° for the right side direction, where the CTC systems exceeds this range. Similar findings can be made for the elevation deviations. For frontal sources a range ±9° is stated in literature and a range of ±10° for sources elevated to 36° in the median plane. Especially for the frontal sources, a wider spread of

results can be observed during the CTC reproduction. Additionally, the findings have to be related to the distance between the virtual source and the nearest loudspeaker of the reproduction array. Especially source position three is very prone to a mismatched participants height for VBAP as the source is on a line between a loudspeaker on the bottom ring of the array and one on the upper ring. That is, a line perpendicular to the horizontal plane. Furthermore, the virtual source's starting position was always the frontal direction. Therefore, the panning distance to the reference source increases with increasing source position number.

For position four, VBAP should excite the same signal as the reference source. Yet, neither median values nor the variation of the results show a decreased deviation. The sound sources provided by VBAP close to the reference position only differ little to the one at the reference position. Additional energy from the other two sources of the active triangle are close to zero and inaudible. Consequently, neighboring source position are accepted as final position.

Last, it should be noted that a coloration of the virtual source (which especially for the CTC system is also position dependent) always leads to a different sound source perception.

4.2.5 Conclusion

The results of the listening experiment show, that the uncertainties in provided localization accuracy between virtual source positions match with those ranges found for natural hearing, especially for the horizontal plane, even though the design of the listening experiment intends to cancel out those effects. While for the horizontal deviations the CTC and VBAP system perform equally good, the results for deviation in elevation show mean values closer to zero for sources reproduced using the CTC system.

4.3 Spatial Audio Quality

While provided localization accuracy and loudness of reproduction system are clearly defined and quantitatively measurable characteristics, more complex terms like naturalness or sound envelopment can only be assessed qualitatively As shown in Section 2.3 the SAQI test method offers a scientific approach towards characteristics specifically selected and formulated for spatial audio reproduction. To gain knowledge about the strengths and weaknesses of the pure reproduction methods, a SAQI test is done. Based on the results, the reproduction systems for the different time segments of the RIR are selected, which are then combined

into a hybrid system a shown in Chapter 6. The results will also be the baseline for evaluating the performance of the hybrid system.

4.3.1 Method

As method, the SAQI test procedure and its qualifier definition and description is used. After answering whether a difference between two stimuli is audible, the participants had to rate the difference for up to four qualifiers on a five point scale (see Section 4.3.2). A comparison against a real or internal reference would be ideal, yet, the set-up used is located in a room with reflections and uses loudspeaker-based reproduction. Furthermore, the differences between the reproduction systems (especially between HOA and VBAP) are expected to be rather small. Three possible ways of using a reference come to mind. The first would be a binaural presentation of the stimuli over headphones. A headphone reproduction would use the same binaural synthesis as the CTC system and requires an individual HpTF equalization and, consequently, would be biased towards the CTC system. The second option would be an internal reference which could be built up in a specific room before the listening test. Considering a listening test with a duration of up to 60 minutes and minimal differences between the presented stimuli, an influence on the internal reference due to the continuing presentation of stimuli is likely and further reduces the measurement accuracy of the test design. As a third option, a comparison against a real source in the room (as used in 4.2 and 4.1) can be considered but would only allow for testing the listening room with its short reverberation time and hence, is not suitable for qualifiers regarding reverberation. As the reproduction systems were tested for suitability in combination with room acoustics simulation, knowledge about their suitability to provide a perception of the surrounding room acoustics in the virtual scene is needed and therefore different rooms with increasing reverberation times were used. Consequently, the use of an internal or external reference was dismissed in favor of a direct comparison of the reproduction systems to allow for higher measurement accuracy and access smaller differences between the reproduction systems. The downside of this approach is, that no direct conclusions towards the reference scene can be made.

4.3.2 Setup

The test was conducted in the VR-Lab with dimmed lights and an illuminated cross in the frontal direction. Participant's positions were aligned as described in Section 3.7.1 on a non-rotatable chair, equipped with the tracking glasses and the tracking system calibrated towards the frontal direction. Interaction with and

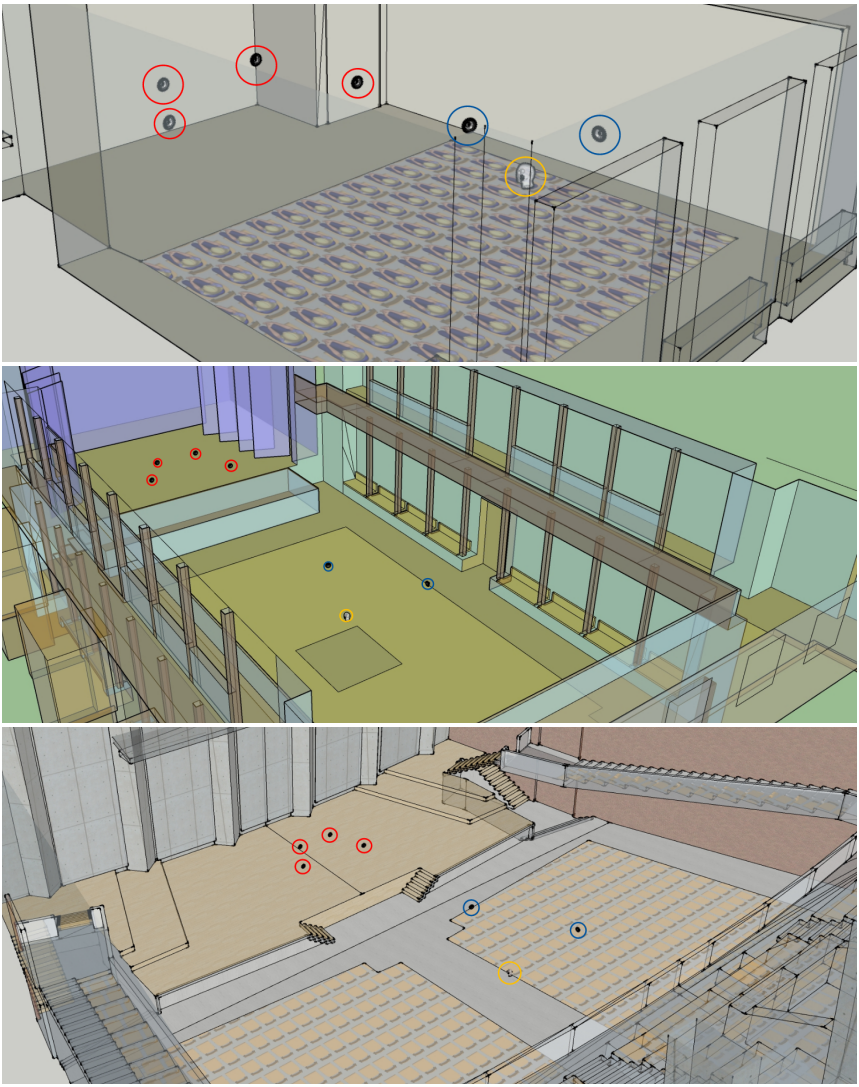


Figure 4.5: The virtual rooms used in the listening test. The upper picture is the small room, the middle one the medium room and the bottom one the large room. The receiver position is indicated with a white head and a yellow circle. The two closer sources on the right-hand side of the receiver are the speech sources and indicated by loudspeaker in a blue circle. The farther located sources are the string quartet from the music scene, indicated by loudspeaker in a red circle. The relative position of the sources to each other do not change in the music scene.

display of the GUI was done via a tablet. During the test, participants were able to freely move their heads. The test started with a training phase (see Section 4.3.2) followed by six blocks in the testing phase. Two stimuli were presented, and the participant had to rate whether an overall difference was audible. If not, the group of qualifiers for this comparison was skipped, and otherwise they had to rate up to four different qualities on a five point scale. For each page, an annotation section was available for feedback or problems. After the testing, an oral interview was done regarding "missing descriptors" (differences that were heard, but were not covered by the perceptual qualities offered), "strategies used" (focusing on certain points in time, association with distinct perceptual qualities), "artifacts" (what kind of artifacts were perceived) and "general problems" (GUI handling, understanding the task or perceptual quality).

Scenes and stimuli

As independent variables, three different virtual rooms were used with increasing reverberation time as the main indicator (see Section 3.7.3). Each room is divided into two scenes with different source positions and different input signals: The first one focuses on speech intelligibility, source geometry and localization (front-back confusion) and utilizes position two (Section 3.7.2) for a female voice presentation and position three for male voice presentation from the former listening test (Section 4.1 and 4.2) at the reverberation distance of each room. The distance was chosen to emphasize the direct sound without suppressing room acoustical effects. This setup will be called "speech scene". As input signals for the speech scene the two anechoic recordings for English language from the Bang & Olufsen CD "Music for Archimedes" (1992) were taken, one with a female voice, the second with the male voice. The scene started with a 1.6 second playback of the female voice, followed by a 1.6 second playback of the male voice. Then, both voices were played back together for about four seconds. This resulted in a seven second dry stimulus as input for the simulation (i.e., rendering) where the reverberation of the specific rooms is added. This approach mixes different independent variables of the listening test: pitch of the speaker, source position and complexity of the scene (i.e., number of active virtual sources) and allows for a more general evaluation of the reproduction systems. As the signals (and the speaker) are unknown to the participants, changes in coloration are expected to be harder to detect.

The second one focuses on the perception of reverberation and room. A fixed string quartet arrangement forming a semicircle arrangement is used (see Figure 4.5). The distance is about three to four times the reverberation radius, hence,

emphasizing the diffuse energy and its decay for a better perception of the room acoustics. The input recordings for the music scene were four second excerpts from the anechoic audio as provided in [Bri+19]. The source positions for this scene, as well as the receiver position, are mainly in the frontal direction, as can be depicted in Figure 4.5. The source positions are taken from the BRAS Database [Bri+21; Asp+20] and sources, as well as receiver, elevated by 0.11m to match the receiver position to the center of the loudspeaker array in the VR-Lab, see Table 4.2 for detailed data on source positions. The instrument recordings have a wider bandwidth and are more easily related to an inner reference. Hence, coloration is expected to be detected more easily.

Source	small room			medium room			large room		
	x	y	z	x	y	z	x	y	z
Pos1	-1.50	1.14	-4.00	-1.50	2.14	-9.86	-1.50	1.14	-10.00
Pos2	-0.76	1.14	-5.2	-0.75	2.14	-11.16	-0.75	1.14	-11.30
Pos3	0	1.61	-4.00	0	2.61	-9.86	0	1.61	-10.00
Pos4	0.75	1.14	-5.20	0.75	2.14	-11.16	0.75	1.14	-11.30

Table 4.2: Source positions used in the music scene of the SAQI listening test relative to a receiver position at [0 1.34 0] meter.

Implementation

The listening test focuses on static sources to avoid additional, variable parameters and computational costs. For the sound reproduction, VA was used in combination with RAVEN. To free up computational power, the stimuli for VBAP and HOA were pre-rendered, as they do not rely on the listener's position or orientation and therefore do not adapt during sound presentation. For the CTC system, on the other hand, a dynamic binaural synthesis as well as dynamic adaption of the CTC filters is needed. Consequently, a combination of RAVEN for the binaural signal and VA for the CTC filters was used to provide (real-time) room acoustics. As the overall calculation time of each RIR is longer than the requirements for a real-time system, the calculation of the DS is prioritized over the calculation of ERs. The calculation of the DD has the least priority. The measured update rates at a block length of 2048 samples and a sampling rate of 44100 Hz can be found in Table 4.3. Each block has a duration of 46.4 ms (2048/44100 Hz) and the DS is updated in every block (for the small room the same is true for ERs). The effect of different block sizes is discussed in Section 6.1.2. The higher block size was chosen to increase filter update rates and reduce mismatch between DS and ERs

Room	Update rate [1/s]			Calculation time		
	DS	ERs	DD	DS	ERs	DD
small	21.5	21.5	0.1	46.4 ms	46.4 ms	10.2 s
medium	21.5	2.73	0.008	46.4 ms	366 ms	125 s
large	21.5	1.7	0.006	46.4 ms	588 ms	167 s

Table 4.3: Calculation times for the binaural synthesis with room acoustics separated in DS, ERs and DD. The test was done with a block size of 2048 samples which calculates to 46.44 ms at a sampling rate of 44100 Hz.

as far as possible. While the perception of changes in the DD due to changes of listener's rotation and position most likely have negligible influence (see Section 2.4.4), the low update rates of the early reflections should be considered when evaluating the results. It also should be noted that the calculation times do not correspond to the latency of the audio stream of the system, but only to update of the filters. Nevertheless, especially the DD needs an initial calculation to avoid zeroed filters. Therefore, each block of the listening test uses only one room and the pauses between blocks were used to initialize the room acoustical simulations. Furthermore, each block started with a comparison of the pre-rendered VBAP and HOA stimuli.

Training

The SAQI provides a description of each perceptual quality for the test conductor (which should be an expert in the field of evaluating virtual auditory environments). The task of how to transport an unambiguous understanding of these descriptors to the participant is left open. To ensure a homogeneous understanding of the perceptual qualities the requirements for participation were set to "to work with acoustics or audio on an at least a weekly basis", which did not include to be working with spatial audio specifically for the sake of gathering enough participants for sufficient data. As a description of the perceptual quality does not necessarily imply practical detection of the intended qualities in stimuli, a training with specific stimuli was conducted at the start of the listening test in addition to providing the descriptors from [Lin14]. To avoid any bias towards the simulation engine, reproduction systems or input signals used in the comparisons, the training stimuli were generated in a digital audio workstation using audio effects. Perceptual qualities like "artifacts" or "naturalness" were not trained, as multiple factors influence them. By artificially degrading the qualities, the participants would be steered towards only single degrading factors.

Qualifiers

The SAQI manual offers 48 qualifiers and their descriptors. As these descriptors are multiplied with the different scenes and comparison of reproduction techniques, they had to be narrowed down to achieve a reasonable testing time. First, qualifiers regarding localization and loudness were eliminated, as they were already tested (see Section 4.2 and 4.1). Secondly, the section "Dynamics" and "Temporal behavior" ("Crispness" was kept) were skipped due to the short stimuli. "Externalization" was skipped due to the fact, that the stimuli were reproduced over loudspeaker. The category "Artifacts" was summed up to just one "Artifacts" and the qualifiers "source width", "source height" and "source depth" to "source extension". The descriptor for artifacts was "Distortion, clipping and not associated sounds of impulsive or tonal character" (less/more pronounced) and for source extension: "Perceived extension of the source in width, height, depth" (less/more extended). The qualifier "Speech intelligibility" was not accessed for the music scene due to a lack of speech.

4.3.3 Results

20 participants were tested with a mean age of 29.9 years ($\sigma = 4.48$). The test duration ranged from 45 to 100 minutes. Figures 4.6 shows the results for the speech and Figure 4.7 for the music scene. The five point rating had two points on each side, and the maximum was rated with double the amount of the midpoint between neutral and maximum. No difference heard was treated as zero value. For each qualifier, the three virtual rooms are color coded (see legend) and show the mean value and standard error as an indicator for the uncertainty of the mean. The value shows a rating towards emphasized/more pronounced/higher in general, "brighter" for tone coloration, "sharper" for sharpness, "more rough" for roughness, "more" for reverberation level, "longer" for reverberation duration, "more distant" for distance, "more extended" for source extension, and "frontal" for the front-back position (labeled "Frontal pos." in Figure 4.6). Data points with a point on the no difference line only indicate that none of the participants indicated a difference heard.

The results for the comparison of HOA and VBAP show a very similar reproduction when using these systems. Only for single combinations of room and scene a difference can be seen. Hence, the results between the comparison of CTC and VBAP and the comparison of CTC and HOA are very similar. The music scene was perceived as brighter in terms of tone coloration when being presented over HOA or VBAP. The perception of higher frequencies for these reproduction methods is consistent with these findings, as well as the tendency

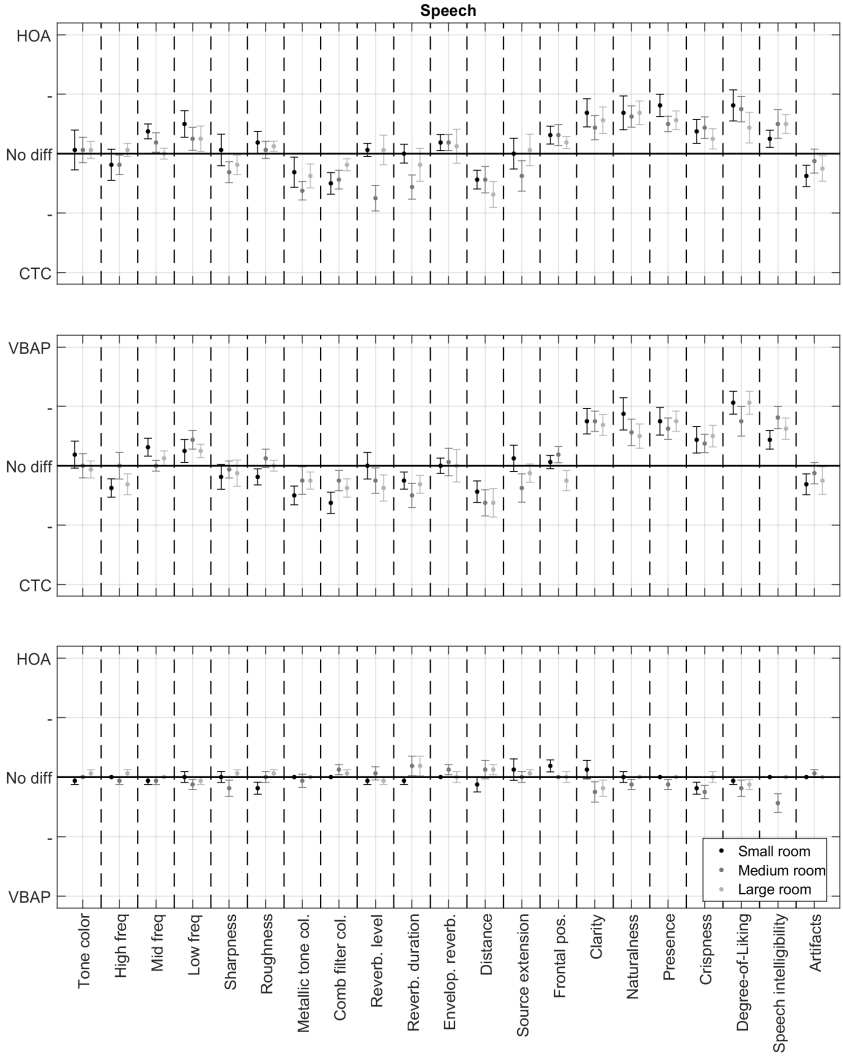


Figure 4.6: Results of the listening experiment in speech scenario as mean value and standard error. Color coded are the different virtual rooms (see legend). The y-axes show the reproduction systems compared. A mean value towards a reproduction system indicates, in general, brighter/more pronounced/emphasized (for details see Section 4.3.3).

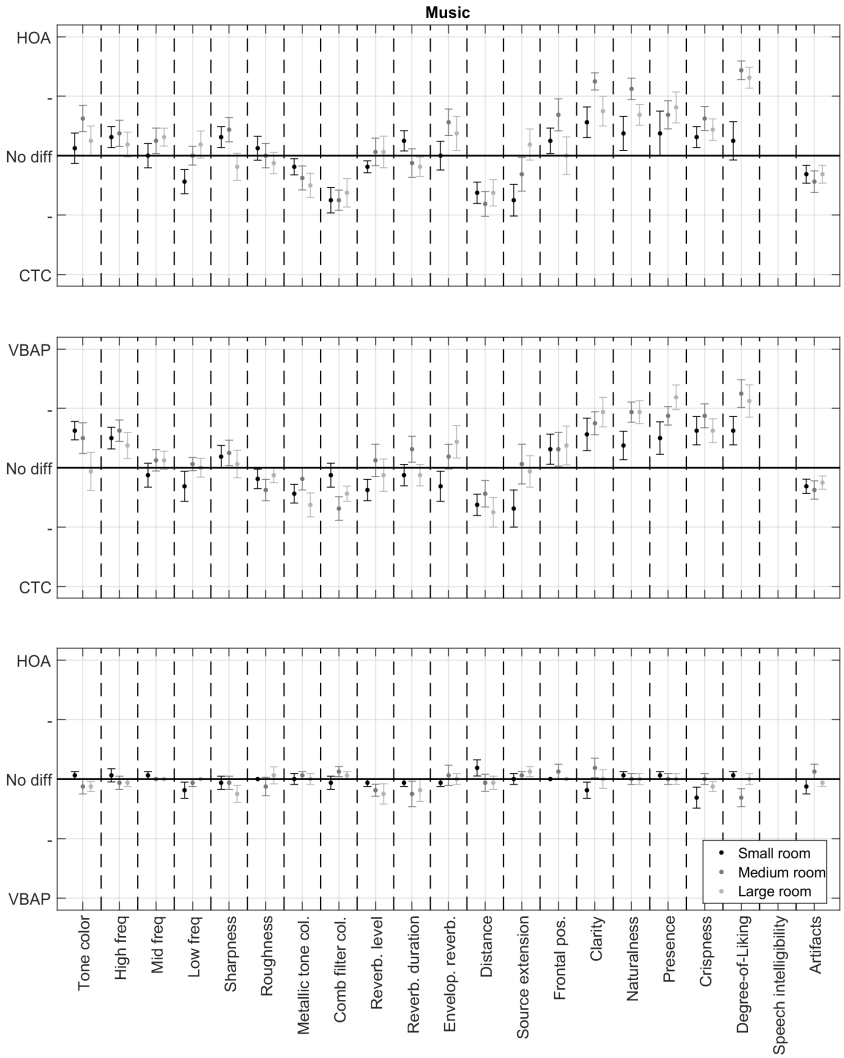


Figure 4.7: Results of the listening experiment in the music scenario as mean value and standard error. Color coded are the different virtual rooms (see legend). The y-axes show the reproduction systems compared. A mean value towards a reproduction system indicates, in general, brighter/more pronounced/emphasized (for details see Section 4.3.3).

for emphasized low frequency perception for the CTC reproduction in the small room. In contrast to the music scene, the low frequency tone color is emphasized for VBAP and HOA reproduction of the speech scene and no difference in tone color can be seen. For the music scene, a tendency can be observed towards a more sharp perception during the HOA (and VBAP) reproduction. Metallic tone coloration as well as comb filter coloration are more pronounced during CTC reproduction of both, music and speech, scenes. The reverberation category does not show significant differences, yet, a tendency for less perception of envelopment by reverberation for CTC reproduction in the music scene but longer reverberation duration for the speech scene. For both scenes, the sources of the CTC reproduction were perceived as farther away. For both scenes, the source extension is perceived partly more extended during CTC reproduction, while the front-back position is more frontal when using VBAP and HOA for the music scene. For both scenes, clarity, naturalness, presence, and crispness as well as degree-of-liking are more pronounced/higher during reproduction over VBAP and HOA. More artifacts were heard when listening to CTC reproduction and in general labeled as inconsistent source positions and in two cases as double source positions present. The oral feedback did not reveal any missing qualifiers or problems in understanding the task. The overall duration was considered too long, as well as finding differences in almost identical stimuli exhausting.

4.3.4 Discussion

While in general VBAP and HOA are based on the similar principles, it was not clear how similar the perception of their qualities is. The results show that almost no differences can be found, similar to recent findings made between VBAP and fifth order HOA in a more densely loudspeaker grid [GHG24]. The results of tone color and low/mid/high-frequency color are linked and will be discussed together. As shown in Section 2.7 and 2.8 the ratio of coherent signals emitted by the loudspeaker may lead to more low frequency coloration. Increasing the room size generally also increases the number of reflections and therefore decorrelates the sound arriving at the listener. Yet, this decorrelation process only applies for sound arriving later than the DS. Moreover, this process of decorrelation only works for reflections within the same order of intensity magnitude. Thus, increasing the source distance and with it the diffuse to direct energy ratio is the second influencing factor to reduce the low frequency coloration. This would explain the difference in the findings between the music and speech scene and the effects in the speech scene. Yet, no reason can be determined for the high-frequency amplification for VBAP and HOA reproduction of the music scene. Furthermore, no well-founded reason can be found for the sharpness results, but the CTC and

binaural synthesis filters might attenuate the frequencies in which the instrument appeared as sharp. The comb filter coloration is expected for CTC reproduction, as shown in Section 2.6. As the CTC with its filter inversion and exchanges is, in general, prone to coloration, the results for the metallic tone coloration might relate to the same indicators as the comb filter coloration. The CTC system also underlies a latency regarding the filter updates of the reproduction, which likely leads to the perception of a more extended source. Both factors, coloration and latency in filter updates are considered as main indicators for the results in clarity, naturalness, presence, crispness, degree-of-liking and speech intelligibility.

The differences in envelopment of reverberation can only be found in the music scene and larger rooms where the reverberant part is emphasized. The hypothesis, that the higher number of active loudspeakers also creates a more enveloping sound perception for HOA reproduction, especially in situations with shorter reverberation times, cannot be confirmed. The complexity of the rooms (and the resulting room acoustics) generate a reflection density for the critical time segment that is high enough to overcome the benefits of a high number of loudspeakers for being active for a single reflection. The difference in distance perception cannot be related to a reference at this point but will be further analyzed in Section 6.2.1. The same is true for the frontal positions. Artifacts reported were mainly towards short appearances of ghost sources and sources moving unexpectedly. This most likely relates to the slower update rates of the ERs during CTC reproduction.

4.3.5 Conclusion

The rather complex listening experiment tested numerous perceptual qualities in two different scenarios for three different rooms with increasing reverberation time. While the setup of the scene does change the perception of distinct perceptual qualities and no obvious influence of the increasing reverberation time was found. VBAP and HOA resulted very similar results in both comparisons: in-between almost no differences were heard and both show the same differences towards the CTC system. VBAP and HOA both benefit from less pronounced comb filter effects and metallic tone color while at the same time providing higher clarity, naturalness, speech intelligibility, crispness, and presence. Two reasons that may partly explain the results are the coloration characteristics of the CTC filters and the fact, that the CTC reproduction is a dynamic system which uses constant filter updates to compensate for the participant's movement. VBAP and HOA, on the other hand, are static reproduction methods that do not compensate for these movements. These differences are also hinted in the reported artifacts.

The assumptions made towards the perception of the room, especially for reverberation envelopment, cannot be confirmed and results are rather inconclusive. The sources for CTC reproduction were perceived further away. No reference source was given during the experiment, and no conclusion can be drawn whether the CTC system is more accurate or not. The test design will be adapted accordingly in Section 6.2.1.

5

Compensation of the listening room

The auralization of virtual rooms aims at reproducing a virtual scene as authentic as possible. Loudspeaker-based reproduction changes the representation by adding the room acoustics of the room where the loudspeakers are placed in (listening room). While free-field environments exist, they are rare, cost intensive and not feasible, neither for wider clinical assessment of patients nor for home entertainment purposes. Different attempts to compensate the room acoustics of the listening room have been proposed [GP15; SBR06; LGF05; TZA14], but mainly focus on the reproduction side.

The methods shown in this chapter utilize the ability to change the virtual room so that the overall reproduction (including the listening room) matches the sound field of the intended scene (i.e., the unchanged virtual room). The two approaches shown are separated into an early reflection compensation (using the results of the image source method) and adapting the slope of the late, diffuse decay.

5.1 Early reflections

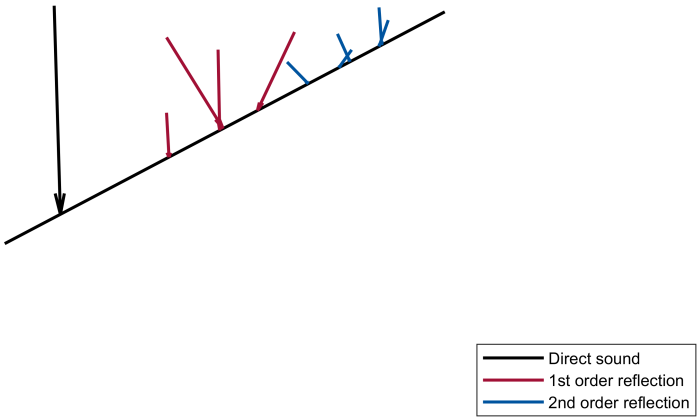


Figure 5.1: Visualization of incoming reflections up to second order for loudspeaker 1 (front left, see Section 3.2) in the listening room. The indicated axis is a time axis and also the frontal direction of the listener. Each tip of an arrow points to the same spatially fixed listener, but at a different time. Each arrow denotes the direction of arrival of the incoming reflection, as well as its time of arrival. The magnitude of the arrays corresponds to the intensity of the reflection. Only the DS is wanted for reproduction of the virtual room.

When reproducing sound over a loudspeaker, the loudspeaker will inevitably excite reflections in the room it is placed in. The main characteristics of these reflections are their intensity and time and direction of arrival, leading to a four-dimensional representation. To illustrate this four-dimensional relation, it can be broken down to a three-dimensional representation, where the time axis is aligned with the x-axis as shown in Figure 5.1. The Figure shows the arriving reflections at a listener position over time. Note that the listener is spatially fixed and only moves along the indicated axis in time so that each arrow points to the listener. The direction of the arrows indicates the incidence angle of the reflections and the magnitude of the intensity at a given time. The intensity of the incoming reflections depends on the directivity of the loudspeaker as well as the room characteristics. These reflections are added to the desired RIR of the virtual room. The proposed approach aims at substituting reflections occurring in the virtual room by using the reflections introduced by the loudspeaker in the listening room due to a reproduced previous reflection. For substitution, prediction of the occurring reflections in the playback room is needed and a "distance" between the virtual reflection and the one in the playback room has to be defined. Prediction of the reflections in the playback room is realized in a hybrid method combining measurement for the intensity and time of arrival of the reflections and simulation for mapping them to the loudspeaker they originated from. For this, a reflection detection algorithm is introduced. The findings in this chapter have also been published in [KV19b].

5.1.1 Listening room - hybrid model

To gain information about time, intensity, and direction of arrival of each early reflection, the listening room will be simulated. The geometrical model should be as accurate as possible. A pure simulation model requires exact input parameters. To describe the suspended ceiling and the curtains in front of the walls, a boundary condition based on complex frequency-dependent data is required. Especially the inhomogeneous air gaps behind the curtains are challenging to acquire regarding their actual reflection factors. To get information about the room as precise as possible within a reasonable amount of effort, a hybrid model was used. A RIR was measured for each loudspeaker with an omnidirectional microphone. The microphone was placed off center to avoid overlapping of single reflections due to the symmetry of the setup. Using the reflection detection (see Section 5.1.2) the virtual model can be adapted and finally used to determine the path of the reflection to extract the direction of arrival of the reflection.

5.1.2 Reflection detection

The reflection detection was developed to determine single ERs in a RIR measured with an omnidirectional microphone to map these on the simulated RIR (with directional information on the incoming reflections) and adapt the (virtual) room model (see Section 5.1.1) to gather information about the direction of arrival. The compensation aims at ERs with sufficient energy to result in an audible change. Based on these requirements, the assumption is made that the desired reflections are similar to the DS in time structure and are reflected by surfaces with limited absorption. The DS is detected in the measured RIR and contains the frequency response of the loudspeaker. The maximum of the RIR is used as starting point. A very short moving window of 5 samples is used to look for the first window with no data above -30 dB to both sides of the maximum. The window has to be short enough to exclude the first reflection after the DS, but long enough to provide sufficient information for a reliable correlation analysis (see below) and not cutting off the window too early due to one or two samples with low amplitudes. The DS is then correlated as a moving window with the RIR and reflections above a threshold are detected as reflections. In contrast to cross correlating the DS with the RIR the correlation coefficient is not biased by the energy in the signal as solely the course of the RIR is analyzed. Yet, this approach has a certain probability of detecting reflections due to random correlation of the noise with the DS. Consequently, the determined DS should not be too short, and the reflection detection should only be performed on a reasonable short time section of the RIR that contains ERs which can be motivated by the mixing time (see 2.4.4). The threshold for detecting a reflection has to be high enough to avoid false detection, yet low enough not to miss out on single reflections. Figure 5.2 shows a normalized RIR measurement and the correlation coefficients. Detected reflections with a correlation coefficient above 0.65 are indicated by circles. In contrast to the RIR the correlation coefficient does not decay as it is not dependent on the energy of the signal. In the late part of the RIR reflections might be detected by chance due to the random course of the noise.

5.1.3 Reflection distance

To achieve a decision criterion whether a reflection can be substituted or not, a reflection distance between a reflection in the virtual room and a reflection in the listening room is defined. As a first approach, the distance is defined by the Euclidean distance d between two n-dimensional vectors:

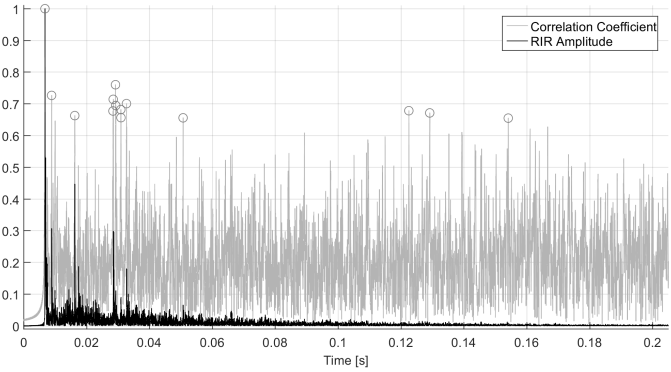


Figure 5.2: Measurement of a RIR with an omnidirectional microphone in black. The correlation of the DS with the signal is illustrated in light gray. Circles denote a detected ER with a correlation higher than 0.65.

$$d = \sum_{k=1}^{dim} \sqrt{(x_{1,k} - x_{2,k})^2} \tag{5.1}$$

The dimension considered are the difference in time of arrival Δt, the direction of arrival $\Delta \overrightarrow{r}$ as well as the intensity of the reflection ΔI. Adding an individual weight w to each dimension equation 5.1 becomes:

$$d = \sqrt{(w_t \cdot \Delta t)^2 + (w_d \cdot \Delta \overrightarrow{r})^2 + (w_I \cdot \Delta I)^2} \tag{5.2}$$

The definition of the individual weights as well as the maximal distance are motivated by perceptual aspects and have to be determined by subjective measurements. It should be noted that each weighting can be further divided into more detailed weights, e.g., the intensity weighting might be time-dependent, the direction of arrival weighted differently for deviations in the horizontal and median plane and the time weighting might depend on the reverberation time as also mentioned in Section 2.4.2.

5.1.4 Implementation

To find suitable reflections for substitution, the simulated RIR has to be convolved with the RIRs of the loudspeaker in the listening room. The convolved, final RIR has to be time-shifted to compensate for the traveling time of the sound from the loudspeaker to the listener for a correct comparison of the RIRs in time (i.e., the arrival of DS has to match). Each substitution of a reflection

in the virtual RIR is a deletion of a reflection which affects the convolved RIR. Therefore, for each reflection substituted, the convolved RIR has to be updated. Before deletion, a check has to be performed to avoid that the closest reflections stems from the virtual reflections to be deleted. Ideally, the deletion of a single reflection would be realized by removing the corresponding entry in the wall hit log of the simulation software and re-generate the filters for the RIR. However, for this proof-of-concept a more efficient approach was chosen by using a time-window with an attenuation of -25 dB to suppress the reflections to be deleted and ensuring a constant transition of the time signal avoiding strong changes in the frequency response. As shown in Section 2.4.2 the hearing threshold of a single reflection is between -20 and -25 dB. Figure 5.3 shows the results for a single loudspeaker reproduction reproducing its own RIR, i.e., the impulse response it creates in the room it is placed in. As the DS excites all reflections, the adapted RIR should be reduced to the direct sound only. In the uncompensated case, each reflection is added by a playback of the reflection itself and the following RIR. In the compensated case, only the minimal energy from the non-perfect deletion of the compensated reflections can be seen.

While this approach works for VBAP and HOA the CTC system has to be approached differently. Additional energy of reflections of the CTC filtered loudspeaker signals should be considered decreasing the minimization of the error made by the CTC (see Section 2.6) and decreasing the provided channel separation. A former approach to including the ERs into the CTC filter resulted in strongly colored signals at the ears without improvement of source localization [Koh+16]. To include the CTC reproduction into the approach and the subjective evaluation (see Section 5.1.5) the BRIR of the virtual room was compared to the BRIR of the listening room including the CTC filter. Unwanted reflections are now processed by the CTC filter, yet, a compensation motivated by the arrival of energy at the listener's is still justified.

5.1.5 Subjective proof-of-concept

The proposed method aims at changing the ERs only. Therefore, the question arises whether these effects are audible or not and whether the produced signals contain any unexpected effects (artifacts). Consequently, a listening test was conducted to investigate the audibility of the compensation and the stability of the signals. As shown in Section 2.4.2, ERs can affect different aspects of perception. Image shifts occur for ERs that a very close in time to the DS or have significantly higher energy. Both cases are very unlikely for typical listening rooms. The same is true for distinct reflections that arrive very late in time.

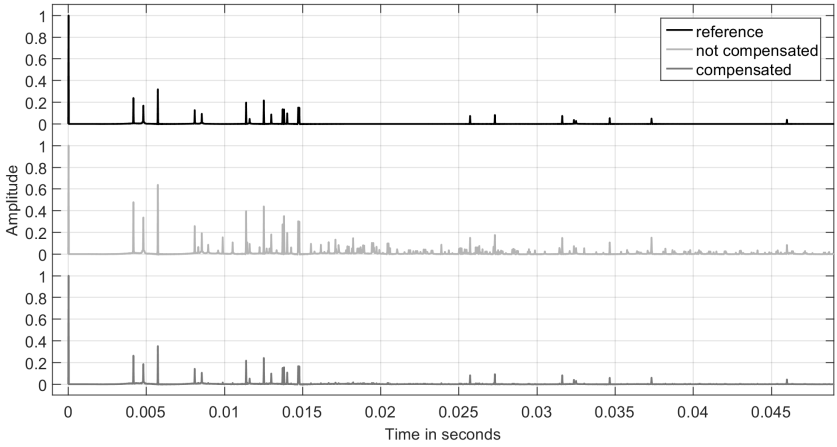


Figure 5.3: Concept demonstration: A single loudspeaker is recreating its own RIR. The black line illustrates the target RIR, in light gray the uncompensated RIR at the listener position is shown (i.e., the target RIR convolved with itself). Dark gray shows the results at the listener position, with the compensated RIR consisting of the DS only and some minor energy in the attenuated reflections.

Testing towards these perceptual qualities would require a kind of artificial setup that most likely would focus on one single reflection and simply test the difference if it either exists or not. The compensation can be expected to change the tone color, yet, a perception of just a difference in tone coloration does not allow conclusions towards a more correct presentation of the intended RIR as most likely any change of the RIR leads to coloration. The apparent source width (ASW) (called "Spatial impression" in Figure 2.4 and [Bar09])) can be used under the assumption, that the initial, virtual sound source is a point source. The perceived ASW of the virtual room is widened by the listening room due to additional, lateral reflections. Compensation of ERs should therefore be perceived as a more narrow ASW. The audibility of artifacts was checked in a follow-up interview.

Method

As a general audible difference between the signals is not sufficient to prove that the algorithm improves the structure of the ERs a parameter has to be found that relies directly on the ERs. As the DS is associated with source localization and the reverberation time is influenced by the DD the perceived ASW was

chosen as a parameter for the test. The decorrelation of the signal due to an increased number of reflections is assumed to widen the perceived ASW, the compensation should therefore be perceived as more narrow. As the proposed compensation is a proof-of-concept, it is not implemented in a real-time system. Yet, listener movement would require a recalculation of the virtual RIR and with it a re-calculation of the compensated RIR. To get stable results and access also small differences, a headphone-based evaluation was made. To include the peculiarities of the loudspeaker array and loudspeaker-based reproduction, the binaural signals were recorded using an artificial head recordings (see Section 3.3) at the sweet-spot of the loudspeaker array rather than rendered by simulation. A 2-AFC experiment was implemented as direct comparison and two virtual rooms and six source positions were altered.

Setup

The listening test was set up in an acoustical treated booth for listening experiments. Open Sennheiser HD650 (Sennheiser, Sennheiser, Wedemark, Germany) headphones were used. For the stimuli, three 300 ms pink noise burst with a 200 ms pause in-between were used and placed as omnidirectional point sources in two different rooms. A model of the VR-Lab and one of a cuboid room with the dimension of 12m x 10m x 4.5m and reverberation time of about 1.5 seconds were used. The source positions are, similar to [Koh+18], located at the following azimuth/elevation angles:

- Position 1: -180°/0°
- Position 2: -120°/0°
- Position 3: -60°/0°
- Position 4: -20°/0°
- Position 5: -20°/30°
- Position 6: -20°/60°

Each stimulus, i.e. combination of position, virtual room and reproduction method, was repeated eight times. Participants were instructed to focus on the ASW and chose which sound was perceived as narrower (A or B). They were explicitly instructed to ignore the effects of coloration. An informal oral interview was done afterward.

Results

A total of 34 normal hearing participants with a mean age of 30.6 ($\sigma = \pm 9.97$) years were tested with a test duration of about 33 minutes. Results can be seen in Figure 5.4. A value of '0' indicates a more narrow perceived sound source during

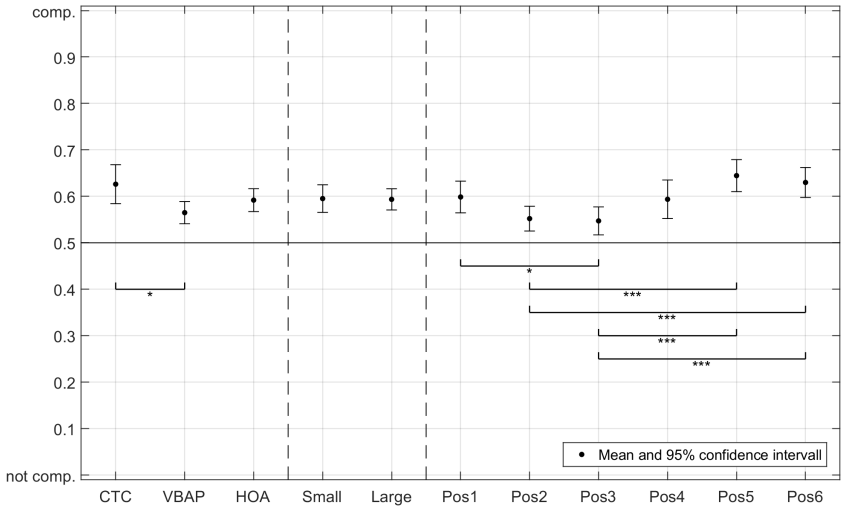


Figure 5.4: Results of the listening test as mean value and 95% confidence interval. Stars indicate significance differences according to their significance level.

the uncompensated reproduction, while '1' indicates a more narrow perception during the compensated reproduction. The x-axis shows the main effects of the listening test, including all combinations that include the listed parameter. Significant differences are indicated with stars according to their significance level evaluated by two-sided t-tests. For all parameters, the compensated playback resulted in a perception of a more narrow ASW. No significant difference can be found for the used virtual rooms. The effect size is slightly larger for the CTC reproduction than for the VBAP reproduction. Position five and six differ significantly from positions two and three, and position three from one. No feedback regarding audible artifacts was given.

Discussion

The results show a measurable effect when using the compensation, even though it might be rather small as the mean values are rather close to the guessing rate of 0.5. The used virtual room does not have a significant influence on the performance of the compensation. While reverberation times (0.15 vs. 1.5 seconds) of the rooms are different, the compensation only affects the ERs and should therefore be more sensitive to different relations between source, receiver and

nearest walls. Position two and three are positioned more to the side than the other ones, yet the effect of the compensation is less audible in these cases, even though the ASW is mainly related to lateral energy. The positioning to the side mainly affects the DS, which is the same in the uncompensated and compensated reproduction, and the floor reflection. The ceiling is acoustically treated and does not deliver any unwanted reflections. For first order reflections from the side walls, a lateral source will mainly result in frontal reflections, while sources at frontal or rear positions mainly evoke lateral reflections. This effect can be seen in the results.

Conclusion

The listening test indicated that the compensation can be applied to the signals without unwanted artifacts, that the difference is audible and can be perceived as a change in the ASW towards the condition of an ideal reproduction situation. It proves that the compensation is not directly linked to the reverberation time but to the ERs and motivates the extra efforts made for the ERs in comparison to the DD (see Section 5.2) as other perceptual parameters are affected.

5.2 Late reverberation

As shown in Section 2.4 the perception of the DD is independent of time of arrival, intensity, or direction of arrival of a *single* reflection. Yet, the overall energy decay can be perceived, mainly as part of perceiving the reverberation (time). As mentioned before: the playback of a virtual target reverberation time is superposed by the reverberation time of the room the loudspeaker array is set up in. An idea proposed in [PV13] is to adapt the absorption coefficients of the surface materials uniformly (multiplied with the same factor) so that the relative difference to each other stays the same. This approach ensures that the timing, direction and relative amplitude of the incoming reflections of the virtual RIR stays the same. The factors needed are approximated iteratively until the target curve is within a predefined tolerance. While this approach may approximate the reverberation time in the DD it cannot change the different curvatures of the decay. To apply the concept to the sound reproduction system, the virtual target RIR is convolved with the RIRs in the listening room. The result is compared to the target RIR and the virtual RIR adapted until the convolved RIRs of virtual room and reproduction array match the reverberation times of the target RIR. To show the procedure of adaptation, a linearly decreasing reverberation time was set as target reverberation time, imitating general absorption characteristics of air and surface materials. Usually, these values would be derived from the target RIR.

The upper plot of Figure 5.5 shows the target RIR as a dashed line and the uncompensated RIR at the listener in solid black. For each iteration, the overall deviation to the target RIR decreases. The initial RIR and the RIR after four iterations can be seen in the bottom part of the figure and plotted over time.

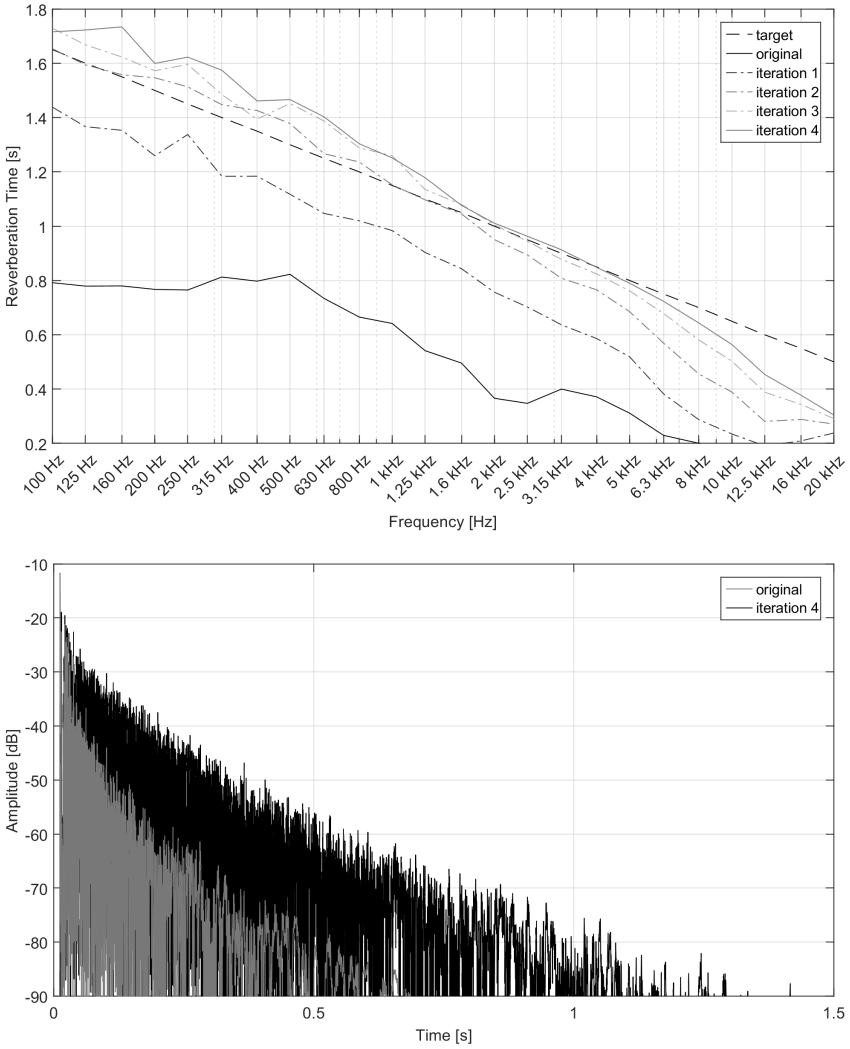


Figure 5.5: Adaption of the energy decay and reverberation time to a fictional target reverberation time. The upper picture shows the adaptation of the reverberation time to a target reverberation time. The bottom plot shows the RIR from the original and the adapted RIR after four iterations.

6

Hybrid system

The initial idea of combining different reproduction techniques was formulated based on their different strengths in reproducing the different perceptual aspects of the different time sections of a (virtual) RIR (see Section 2.4.4). While there are benefits to using CTC for the DS to improve source localization (see Section 4.2) the better perception of envelopment by sound due to a (lower order) Ambisonics was not proven in the SAQI listening test (see Section 4.3). Yet, especially when combined with real-time room acoustics, the CTC system is affected by the constant filter updates and, in the case of static sources, is more prone to delayed filter updates from the room acoustical simulation. Therefore, the expectations towards a combined system based on the results of the SAQI listening test are now formulated in a better perception of clarity, naturalness, presence, and crispness while maintaining the localization accuracy of the CTC system.

6.1 Implementation

For the selection of a suitable reproduction technique for presenting the DS, the results of Section 4.2 are used. The missing data on the HOA reproduction system (see Section 4.2.3 are contextualized by a comparison of the expected performance in relation to the VBAP system. As indicated in Section 2.8 both system are closely related to each other, especially in the specific implementation of HOA used for the hybrid system. Furthermore, VBAP "activates the smallest possible number of loudspeakers" [ZF19] for 3-D audio reproduction, resulting in a directional presentation of the sound source, yet varying with the position relative to the active triangle. This theoretical approach is supported by the findings from the SAQI listening test in Section 4.3 where both systems result in the same differences towards the CTC system and almost no differences between each other. It is therefore valid to expect that the localization accuracy provided by the HOA reproduction is not significantly better than the one provided by the VBAP system and, from a theoretical standpoint, is more likely to be less accurate than the VBAP system. While the results of Section 4.2 show similar accuracies between the CTC and VBAP system for the horizontal plane, the CTC system results in a more precise perception of the source location in elevation on average. Therefore, the CTC is used in the hybrid system to present the DS.

For the ERs both systems, VBAP and HOA, can be considered for the above-mentioned beneficial qualities (coloration, clarity, naturalness, presence, and crispness) compared to CTC system. Weak tendencies in terms of more pronounced degree-of-liking, clarity and crispness in both scenes of the SAQI test and better speech intelligibility for the speech scene are the reasons for the choice

of VBAP as reproduction system for the ERs.

While the comparison between VBAP and HOA in Section 4.3 is rather inconclusive, HOA was chosen for the DD to prove the concept of combining *different* reproduction methods and analyze the effects of the combining concept. A more practical motivation can be found in the easy rotation of the sound field when using HOA, making recalculation of RIR of the computational costly Ray Tracing process obsolete in some cases. Finally, the abbreviation for the hybrid system is set to CVH to name the reproduction techniques in chronological order of the RIR.

As the provided qualities between VBAP and HOA reproduction are very similar, the significance of the transition time becomes less important. As shown in Section 1.1 on average an image source order of two is sufficient. To reduce the computational load and decrease the filter update latency, this image source order of two was chosen rather than the number of three, which is calculated as a conservative number in [PSV11]. Furthermore, the chosen image source order of two complies with the SAQI test in Section 4.3. This allows to reference the results for the subjective evaluation of the hybrid system with the former tests. It should be noted that the image source order is an adjustable parameter for the room acoustics simulation rather than an inherent part of the reproduction system and can therefore easily be changed.

The requirement for a seamless combination of reproduction systems is an inaudible transition between them while maintaining the structure of the target RIR. Two main parameters have to be adapted: loudness and latency. The loudness adaptation is based on the findings in Section 4.1 and shown in Section 6.1.1 while the latency adaptation is laid out theoretically in Section 6.1.2, measured in Section 6.1.2 and transition from ERs to DD subjectively evaluated in Section 6.2.2. In the current configuration, the VBAP filter for reproducing the ERs and the HOA filter for the DD are pre-calculated which, in reverse conclusion, demands static source positions while the user can still move in the loudspeaker array to some extent. The room compensation methods require prior knowledge of the room in which the loudspeaker array is set up in. While the DD compensation part (see Section 5.2) only requires the energy decay of the listening room (i.e., an omnidirectional RIR measurement or simulation) the compensation of the ERs requires a detailed model of the listening room with correct timing and direction of arrival of the incoming reflections to determine the intensity of the reflections.

6.1.1 Loudness

To match the loudness in-between reproduction systems, and to calibrate them to the reference source, the findings of the listening test in Section 4.1 are used in a way, that the mean of the mean values is used as a correction factor. The mean values over all participants can be found in Table 6.1. The correction

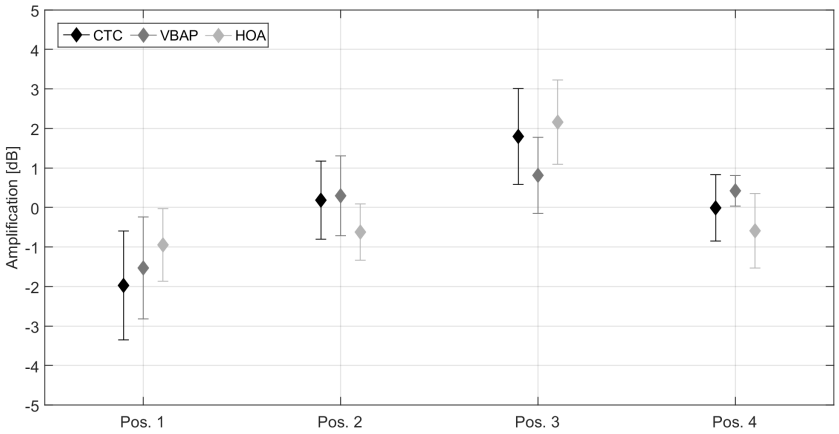


Figure 6.1: Estimated loudness perception of the corrected reproduction systems as deviation to a reference source.

Method	Pos. 1 mean, [dB]	Pos. 2 mean, [dB]	Pos. 3 mean, [dB]	Pos. 4 mean, [dB]	Correction [dB]
CTC	-2.76	-0.60	1.01	-0.79	+0.78
VBAP	-1.69	0.14	0.66	0.27	+0.16
HOA	-0.13	0.20	2.98	0.23	-0.82

Table 6.1: Mean gains needed for participants to perceive the same loudness between reproduction method and reference source in dB.

factor calculates to the negative mean over all positions for a single reproduction method. Under the assumptions of a linear, time-invariant system, the resulting deviations between reproduction system and reference source adjust to Figure 6.1 (each reproduction system is shifted by the correction gain compared to Figure 4.2). The gain applied is a system-specific gain and independent of the virtual source position, hence, the differences between source positions in a single

reproduction method remain the same, yet, the differences *between* reproduction methods in each single position is minimized.

6.1.2 Latency

The latency of the overall system can be considered as reaction time to an event that changes the acoustical scene and divides into two main latencies: The streaming latency, which delays the sound signal from feeding into the system until it arrives at the listeners ears, and the filter update latency, i.e., the time needed to correct the signal at the listeners ears due to a change of the virtual scene, which is typically due to a change of source or listener position or orientation. Figure 6.2 shows the principle of the different latencies. In the case of a CTC reproduction, the filter update latency also includes the update of the CTC filters due to listener movement in the loudspeaker array. The streaming latency can be separated into different aspects: The acoustic latency, which is the time needed for the sound waves to travel from the loudspeaker to the listener's ears (and therefore depends on the loudspeaker distance) and the listener's position. The electrical latency is introduced by the used hardware for digital to analogue conversion and back. The framework latency contains the processing of the data and especially filter convolution. It depends on the available processing (computational) capacity and the used buffer size. While with increasing block size the computational load decreases, it increases the latency with which the new filters can be provided (see below), as new filters cannot be applied during a block. For a high overall latency, the block size becomes less important and can be increased to reduce the latency due to calculation times. Setup-specific values for the streaming latency can be found in Section 6.1.2.

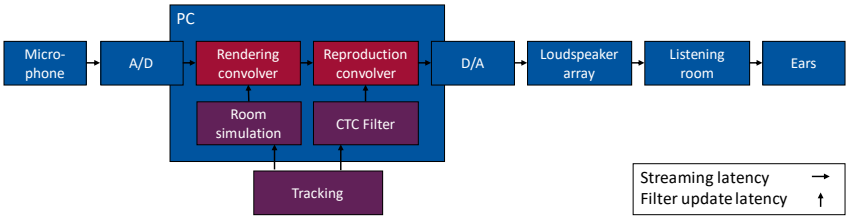


Figure 6.2: Schematics of streaming and framework latency. The microphone and analog to digital conversion are only needed for user interaction. Otherwise, the signals typically stem from a file stored on the PC. Magenta colored blocks show the influence on the filter update rate. Red blocks show the convolution of the running audio stream with the current filters.

On the reproduction side, the streaming latency is mainly defined by the buffer needed to avoid non-causalities in the CTC filter (see Section 2.6). As mentioned before, the filter update latency on the reproduction side is defined by the time needed for the calculation of the CTC filter and is further affected by the tracking system. The tracking system operates at 120 Hz, which results in a maximum delay of 8.3 ms. The tracking also influences the rendering filter update latency as it positions the listener in the virtual room, which then triggers a filter update of the binaural filter. Additionally, the filter update latency on the rendering side heavily depends on the variables in the virtual acoustics scene, like the room size and complexity and the corresponding number of image sources to be calculated and minimum number of rays needed. Additional virtual sources will increase the demands for the computational effort. The required filter from the rendering are room acoustical impulse responses (IRs) and are calculated separately for DS, ERs and DD. In general, the DS should be calculated with the highest priority, followed by ERs and the DD with the least priority due to its position independent perception (see also Section 2.4.4). The rendering latency is a more general problem and not a specific one for the combination of reproduction methods. The influence of the used block size can be seen in Table 6.2.

The filters delivered by the room acoustical simulations are direct loudspeaker signals for VBAP. For HOA the signals have to be decoded. While the decoding matrix may need some calculation power it only has to be determined once, preferably during startup, i.e., before using the system. Once the decoding matrix is available, the decoding consists of simple multiplication and summation of the channels. It can be noted that in general the aspects of streaming latency add upon each other while the filter update latency are mainly defined by the highest occurring value, yet each mismatched filter affects the perception differently. E.g., increased coloration for mismatched CTC filter versus false source localization from mismatched binaural filters.

The main aspect of this section is therefore the latency *differences between* the reproduction systems that occur in the audio stream, rather than those of the filter update rates. VBAP and HOA have a fixed latency that is, in general, smaller than the one of the CTC system which needs a buffer for avoiding non-causalities of the filter inversion (see Section 2.6). As shown in Section 3.4 the latency in the sweet spot is 1024 samples and therefore both signals, of VBAP and HOA, are delayed by 1024 samples (23.2 ms). Small mismatches in latency due to listener movement are not compensated.

Block	All	Small room		Medium room		Large room	
size	DS[ms]	ERs[ms]	DD[s]	ERs[ms]	DD[s]	ERs[ms]	DD[s]
512	11.6	44.3	17.5	538	208	862	331
1024	23.2	44.1	14.3	435	162	690	247
2048	46.4	46.4	10.2	366	125	588	167

Table 6.2: Simulation times in milliseconds for different block sizes, calculating the BRIR of four sources divided into ERs and DD. The direct sound is always prioritized so that an update in every block is possible. Consequently, the update time for the DS calculates to block size divided by 44100 Hz.

Latency measurement

The filter update rates of the system rely on input variables and are not constant. To measure a baseline latency, the filter update rates are avoided by utilizing the configuration from the user integration listening experiment (see Section 6.2.2). In the experiment, participants were able to speak into a virtual room, thus, avoiding the playback of DS over the CTC system. Using pre-calculated filters, the setup only has the streaming latency with the electrical, acoustic and framework latency. The latency is measured using a microphone as input signal. An impulse response measurement was done using a sweep generated by the loudspeaker 1 (front left) and the microphone of the listening test taped to a KE4 microphone capsule (KE4, Sennheiser, Wedemark, Germany) (see Figure 6.3), positioned in the center of the array. The additional microphone was used to feed the signal into a different hardware channel for the measurement software.

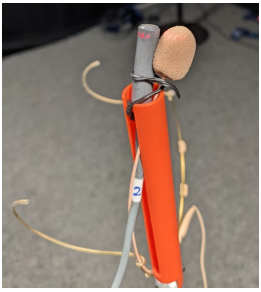


Figure 6.3: Set up of the latency measurement. The microphone used for the listening test was taped to a reference microphone.

The microphone input was then routed to the system, which played back the signal, in the same way as speaking into the room would. The overall streaming latency

Block	Overall latency	w/o leading zeros	Framework
256	20.6	14.1	13.9
512	32.2	25.7	19.0

Table 6.3: Different latency values for the framework latency of the system.

is then the time between the first impulse arriving at the two microphones and the system's response again arriving at the microphones. The streaming latency includes the acoustics transfer path from loudspeaker to listener. At a loudspeaker distance of 2.3 meters and a speed of sound of 343 m/s this latency calculates to 6.7 milliseconds. The system was set up with a VBAP RIR simulated in RAVEN in a total absorbing surrounding (i.e., free field conditions), and the latency of this RIR deducted from the measured latency. The system worked stable at 512 samples buffer size when used in the listening experiment set-up of Section 6.2.2 and at 256 samples without MATLAB and the web interface running. Everything was set up on a single desktop computer (Intel Core i7-4790, 3.6 GH, see Section 3.1). For the system running with 256 samples buffer size, the latency is measured to 20.6 ms. Deducting the 6.7 ms for the acoustic path, the framework latency calculates to 13.9 ms. Assuming a listener height of 1.34 meters and a source (and reflection) free loudspeaker array, cutting the leading zeros of the RIR (6.5 ms) leads to an overall system latency of the 14.1 ms and 7.4 ms without acoustic transfer path. An overview of the measured latencies can be found in 6.3. Using 512 samples of buffer size, the overall system latency changes to 32.2 ms and 25.7 ms when cutting the leading zeros, which results in 19 ms of framework latency without the acoustic transfer path. The latency measurement shows that the system in this basic configuration can deliver sound in real-time (less than 50 ms) and the interaction with the system is possible with live input (less than 15 ms) when cutting leading zeros of the RIR. For the latency between virtual source and listener, the additional 23.2 ms for the CTC filters have to be incorporated. Yet, for delays towards a visual reference higher latencies up to 50 ms are, in general, accepted [Vor20].

6.2 Subjective evaluation

To evaluate the overall performance of the system, a subjective evaluation is needed. The perceptual qualities are tested against the results of the SAQI listening test (see Section 4.3) which served as a reference. The test was further enhanced by a section to compare the provided localization accuracy against a loudspeaker as visual reference and a section to evaluate the effect size of the room

compensation. The second listening test was conducted to test the integration of the user and the interactive capabilities of the system and estimate the need for updating the ERs filter due to listener movement and estimate the required precision in time for the transition between ERs and DD.

6.2.1 Spatial Audio Quality Inventory

To evaluate whether the hybrid system is capable of utilizing the strengths of the different reproduction methods in an audible effect size while not increasing the weaknesses of the systems, the SAQI test design and framework from Section 4.3 is used. Using the same test procedure benefits from a comparability between the results (as is intended by the SAQI, see [Lin+14]). The design was extended to evaluate the localization performance in relation to a visual reference source.

Method

The same method as for the first SAQI comparison was used (see Section 4.3.1), where participants first had to indicate whether an overall difference between two reproduction methods, in this case CVH and CTC (see below), is audible or not. If audible, they had to rate the difference between them. The test was extended by a localization part to evaluate whether the localization accuracy against a reference is degraded by the combination of systems or if it does not lead to significant differences when compared to the CTC system which could be expected as both system share the same DS presentation. For the perceptual qualities regarding reverberation, the CVH was compared with and without room compensation. As the overall testing time in the previous SAQI extended a reasonable duration, the number of stimuli and reproduction methods to be compared were reduced (see below).

Setup

The setup is mainly the same as in the first SAQI test in Section 4.3. One issue of the first SAQI listening test was its testing duration for the participants, which is now extended by the localization part and the room compensation comparison. To reduce the overall duration for the participants, the comparison is only done against one reproduction system. As the CTC system is prone to filter update rates and long calculation times were needed for the DD, especially in the large room the reproduction was optimized by using pre-calculated, static filters for the DD part of the binaural synthesis and thus freeing up calculation capacity for the calculation of ERs. This reduced the filter update time to 46.4 ms for the small room, 303 ms for the medium room and 467 ms for the large room. As

no correlation between perceptual quality and room (or reverberation time) was visible, for each perceptual quality the virtual scene with the highest deviation between VBAP and HOA towards CTC was used to further decrease the number of stimuli which are shown in Table 6.4. The localization part used the source positions 2 and 3 of former listening tests (Section 4.1 and 4.2) with the two white loudspeakers as visual reference. Consequently, the directions of these two sources were identically with the speech scene, yet, the distance was shorter for the localization part. These source positions were then auralized in all three virtual rooms and their localization accuracy was determined in vertical and horizontal direction, as well as in depth (i.e., distance) by asking which stimuli provided source positions closer to the visual reference. A list of the selected qualities and the virtual room they were tested in can be found in Table 6.4. As the CVH system uses a compensation of the listening room, the perceptual qualities regarding reverberation were additionally compared between the CVH system and a CVH system without any room compensation. The same oral interview as for the former SAQI test was conducted after the listening test.

Results

20 persons with a mean age of 28.8 years ($\sigma = 4.05$) participated in the listening test with a duration of about 45 minutes. The results can be seen in Figure 6.4 as mean values and standard error for each perceptual quality and scene, with the latter one separated by color. The y-axis denotes the hybrid CVH system and the CTC system. A mean value pointing towards one of the system in general indicates "more" and "brighter" for tone color (see Table 6.4 for exact terminology). The results are divided into three sections. While the first section sums up perceptual qualities indicating a positive effect, the last one shows those with a negative effect. Due to the lack of a real reference, the middle section can not be rated in a sense of positive or negative effect right away but evaluate whether an audible difference is measurable.

In the first section from the left, the CVH system is preferred by the participants in at least one scene with the exception for crispness where results are less significant. For "tone color" the speech scene is perceived as brighter for the CTC system with low frequencies perceived as "emphasized" during playback over the CVH system. Contrary, the music scene is perceived as brighter during the CVH playback and with a weaker indication for an emphasized perception of the high frequencies part during CVH reproduction. The perception of reverberation shows a longer reverberation time, more reverberation level and a tendency towards more envelopment of reverberation during the speech scene

Perceptual quality	Scene	Virtual room	Marker
Tone color	both	small	brighter
High frequencies	both	small	emphasized
Mid-frequencies	both	small	emphasized
Low frequencies	both	small	emphasized
Sharpness	both	medium	sharper
Roughness	both	medium	more rough
Metallic tone color	both	large	more pronounced
Comb filter coloration	both	medium	more pronounced
Clarity	both	medium	more pronounced
Crispness	both	medium	more pronounced
Naturalness	both	medium	higher
Presence	both	medium	higher
Reverberation level	both	medium	more
Reverberation time	both	medium	longer
Envelopment by reverberation	both	medium	more pronounced
Degree-of-liking	both	medium	higher
Artifacts	both	small	more pronounced
Source extension	music	small	more extended
Source extension	speech	medium	more extended
Speech intelligibility	speech	medium	higher
Source distance	— Tested in localization part —		
Front-back position	— Tested in localization part —		

Table 6.4: Perceptual qualities selected for the evaluation of the hybrid CVH system and the virtual room used. If not further indicated, the set-up was used for both scenes: music and speech. The marker section indicates the meaning of a mean value towards a system in Figure 6.4.

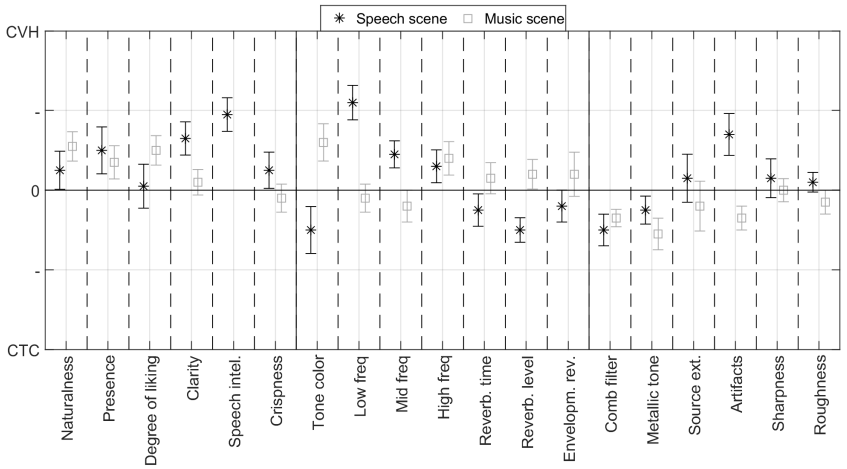


Figure 6.4: Results of the SAQI evaluation of the hybrid system. Markers represent mean values and whiskers the standard error. The y-axis indicates the reproduction system. A marker towards one system generally means "more" (pronounced/emphasized) and "brighter" for tone color. Details of the marker indication can be found in Table 6.4. Different colors separate the music and speech scene.

when using the CTC system. Tendency towards a contrary behavior can be seen in the music scene, where the effects tend towards the CVH system. For the perceptual qualities "comb filter coloration" and "metallic tone color" the CTC results in more emphasized coloration, while for "source extension", "sharpness" and "roughness" no differences can be found. Finally, the speech scene is to be more prone to artifacts when played back over the CVH system, while the opposite is true for the playback of the music scene. The perceived localization is more accurate for both scenes and all virtual rooms when using the CVH system, as shown in the upper part of Figure 6.5. No audible effects of compensating the listening room can be found as shown in Figure 6.5, bottom plot, except for a less perceived reverberation level when using the room compensation. In the oral feedback, three participants rated the difficulty of the test as "easy", 13 as "medium" and four as "hard". No missing descriptors were mentioned. Four participants mentioned a (partial) front-back confusion, from which one stated "only in high frequencies" and one said "for a sharp ’s’ sound".

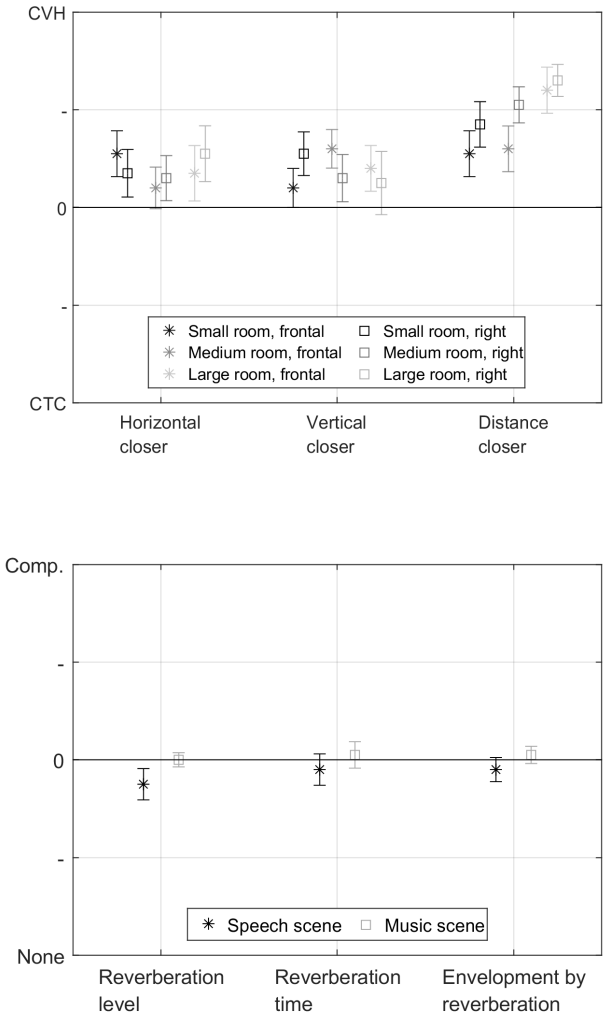


Figure 6.5: Upper plot: Results of the comparison of provided localization accuracy between the hybrid CVH system and the CTC system against a visual reference. Values are stated as mean values and standard error. Values towards a reproduction system indicate a perceived virtual source closer to the reference. Bottom: Comparison of two CVH systems: One with room compensation labeled "Comp.", the other one without labeled "None". Markers indicate "more" perceived reverberation level, "longer" reverberation times and "more pronounced" envelopment by reverberation.

Discussion

During the oral feedback, three out of the 20 participants reported a perception related to a blurry source presentation due to head movement, especially during the "speech scene". This perception might be due to the constant updates of ERs filters during CTC presentation and maybe an insufficient ERs update rate. Both effects were not consistently perceived but only in a "few" stimuli. Furthermore, the evaluation of reverberation was considered difficult due to the length of the stimuli that was between the two reverberant decays (5 out of 20).

To discuss the perceptual qualities, a short recap of the scenes is done: The speech scene uses closer (less distance to the listener) sources with recorded speech as input signals. As the original voice of the speaker is not familiarized to the participant, an absolute estimate of change in tone coloration is more difficult. In contrast, the music scene utilizes instruments, that do have a distinct sound character and can be linked to an inner reference. Furthermore, the sound sources in the music scene are placed far behind the reverberation radius and thus, these scenes focus more on reverberation and systematic coloration rather than on sound localization, separation of sources and clarity as provided in the speech scene.

In the first section, the perceptual qualities "Naturalness" and "Degree of liking" are both linked to tone coloration and both show tendencies for a stronger effect in the music scene. "Clarity" in the context of the SAQI test is circumscribed by "Impression of how clearly different elements in a scene can be distinguished from each other, how well various properties of individual scene elements can be detected" [Lin14]. Together with "Presence" it is more linked to source localization and separation and therefore shows stronger effects for the speech scene with less distant source positions. The speech scene is perceived as "brighter" when played back over the CTC system, and results indicate that this stems from a low frequency boost during CVH playback. This low frequencies boost might stem from the higher direct energy to reverberant ration and the playback of coherent signals over all loudspeakers. Interestingly, the music scene reveal are contradicting behavior, where the CVH system is perceived as brighter. This effect is either dominated by the low frequency boost in the speech scene or more audible as the music signals uses a wider frequency range. These findings comply with those of Section 4.3.3 where VBAP and HOA show similar differences towards the CTC system.

To evaluate the results of perceptual qualities regarding reverberation characteristics, the influence of the room compensation is discussed first. Both approaches of compensating the listening room, for the ERs (see Section 5.1) and the DD (see Section 5.2), decrease the amount of overall energy in the IRs. Yet, it is unclear whether the amount of energy missing is sufficient to be audible, especially as the listening room already has a short reverberation time. The bottom plot of Figure 6.5 shows the result of a direct comparison between a CVH system with compensation (labeled "Comp.") and a CVH system without compensation. No significant differences can be found, except for a tendency towards less perceived reverberation level during the uncompensated playback in the speech scene. The difference between CTC and CVH in Figure 6.4 regarding the reverberation time points towards a dependency on the scene used. While the music scene does not provide significant results, the speech scene does. One reason might be originated in the significant differences found for "distance", another that the filter updates for the ERs of the CTC system create a more diffuse perception of the reverberation. As the music is more diffuse as such, that effect diminishes. The results of the oral feedback also indicate that for the music scene no energy decay can be noticed except for the last decay. This last decay might have been too far apart in time between the two stimuli. Similar to the results for the room compensation, the strongest effect can be seen for the reverberation level in the speech scene, which most likely originates from the listening room compensation. The third section indicates a clear tendency towards the perception of more pronounced comb filter coloration and a more pronounced metallic tone color, without clear indication of a scene dependency. For "sharpness" and "roughness" no significant differences can be found. The inconclusive results for "Source extension" indicate, that the compensation of the ERs does not have an audible effect in more complex scenarios.

"Artifacts" were tested in the small room only. In the speech scene, the artifacts relate to a partial front-back confusion with a high frequency (sharp 's') sound that might stem from an inconsistent localization between DS and a strong ERs. This might be a hint that in this case the inconsistency between the reproduction methods is audible. For the music scene, the same reasons as for the first SAQI test were seen and can mainly be mapped to the filter update rates.

The results for the perceived localization accuracy are shown in the upper plot of Figure 6.5. For all virtual scenes, the CVH playback was perceived closer to the visual loudspeaker than over the CTC system in all three degrees of freedom. No systematic offset between source location (frontal or right) can be seen, neither

a systematic influence of the room size, except for distance, where the deviations between CVH and CTC system increase. Contrary to the expectations, the localization accuracy increased instead of being the same at best.

Conclusion

The listening test indicated that the hybrid system exceeds the expectations for the provided source localization accuracy, as it not only provides the same localization accuracy as the pure CTC system but improves it in all three, translational, degrees of freedom. It delivers a higher "Naturalness" and "Presence" and a higher "Speech intelligibility" and, depending on the scene, can improve "Clarity" and a general "Degree of liking". At the same time, the system benefits from decreasing the CTC typical "Comb filter coloration" and "Metallic tone color". No effects of listening room compensation can be found, neither in a direct comparison nor in a difference in "Source extension". Yet, the listening room used already has a short reverberation time, which leads to smaller energy changes in the compensation.

6.2.2 User integration

To integrate a user into a virtual scene, the self-perception of the listener in the room has to be correct. When speaking, the user generates a direct sound and is thus expecting ERs and a DD as response. This "reaction" of the virtual environment is needed to increase a sense of presence, naturalness and, thus, immersion. A listening experiment with a focus on ERs was therefore designed to test whether such an integration is feasible. During the experiment, participants had to speak into a microphone and the reproduction system responded with a "virtual room", i.e., ERs and DD. The experiment is further designed to find a threshold for audibility of listener displacement in the virtual scene to estimate the requirements of the filter update rates of the ERs. If a certain amount of listener displacement is inaudible, lower update rates can be realized, which decreases the computational load. The displacement of the listener in the virtual scene only affects the ERs and, thus, offers an insight to the demanded accuracy in timing between ERs and DD, in general and specifically for combining the reproduction methods. As the DS is provided by the user itself, the dynamic CTC system is not used and, due to pre-calculated RIRs, no filter updates are necessary in this setup (see also Section 6.1.2). Consequently, this setup works solely on the streaming latency of the system, i.e., the minimum latency without any filter updates. The results are also published in [PKV23].

Method

To estimate the threshold for a minimum audible displacement a 3-AFC with hidden reference was implemented, with a one up, two down scheme (see below). Participants were presented three stimuli (i.e., RIR) of which two were identical. In the initial phase, the displacement was decreased until one of the identical stimuli was selected, i.e., the differing stimulus could not be detected. In the following measurement phase, they had to identify the different stimulus two times in a row to lower the mismatch and to decrease the influence of identification by chance. The mismatch was created by displacing the virtual receiver in the simulation in either translation or rotation. To reference the perceived ERs in time, the own voice was used as DS and the filters for the DD were position-independent, i.e., the same for all displacements, to focus the conclusions of this test on the ERs only. People were therefore asked to speak into the room and were presented with the reflections of the virtual room over the loudspeaker array.

Setup

When speaking into a room, the receiver and source position are identical in the room simulation model. The mismatch was created by moving or rotating the *receiver* to simulate a mismatched listener in the loudspeaker array. The choice of displacing the receiver (instead of the source) has a direct effect on the rotational displacement: rotating the source only changes the intensity of the reflections due to the source directivity whereas rotating the receiver rotates the entire virtual room around the participant and, consequently, changes the direction of the incoming reflections. The translational displacement was orientated to the frontal direction ($\varphi = 0°$ azimuth) or towards the left side ($\varphi = 90°$ azimuth) in the horizontal plane. Rotation was done around the y-axis, decreasing the azimuth angle of the view vector in the horizontal plane going from frontal direction to direct right. As the test is designed to converge to *one* threshold, the maximum rotational mismatch was set to -90°, where ITD and ILD reach their maximum. For absolute angles higher than 90°, an increased difference cannot be assumed, as ITD and ILD decrease again. The positioning of the user in the virtual room is set, so that the direction of the first reflection, after the ground reflection, is shifting from direct right to the front. This further ensures that increasing the mismatch leads to an increased audibility of the error.

The rooms as described in Section 3.7.3 were used and can be seen in Figure 6.6 together with the receiver position. To increase the audibility of the rather small changes, the source, and receiver were positioned close to a wall on the stages for medium and large room and on one side of the small room as shown in Figure

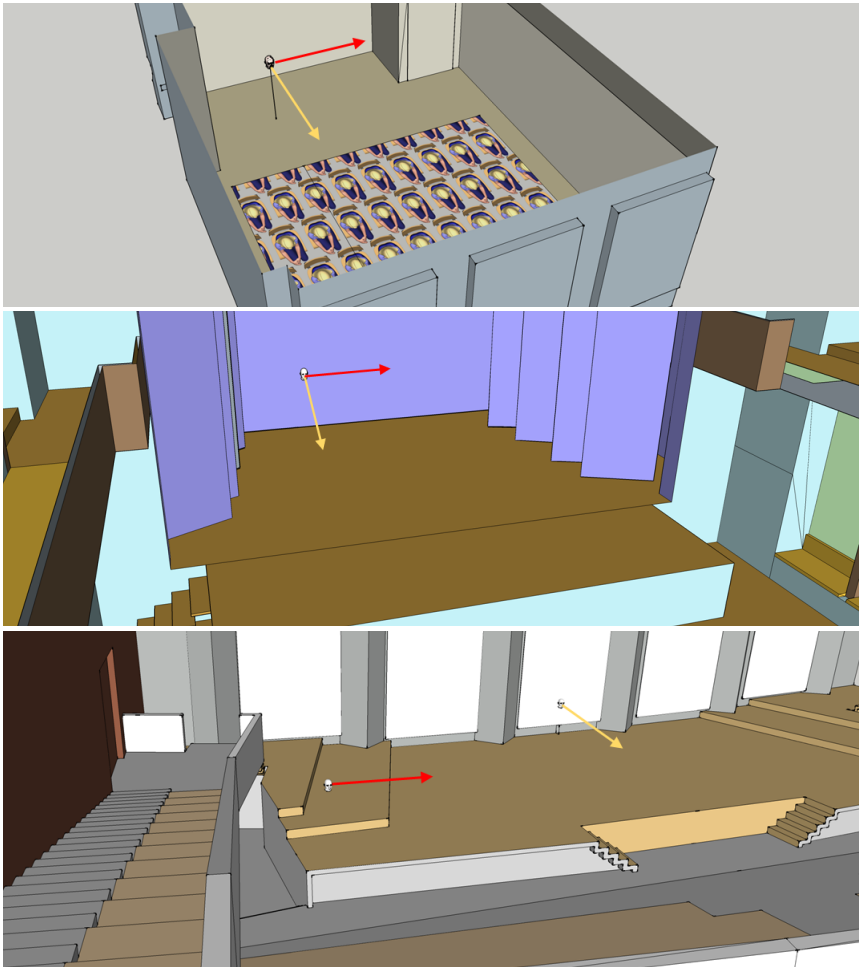


Figure 6.6: The virtual rooms used in the listening test. The upper picture is the small room, the middle one the medium room and the bottom one the large room. The receiver position is indicated with a white head. Frontal displacement is indicated by a yellow arrow, translational mismatch to the left by a red arrow. In the large room, rotation was done at the head position indicated with the red arrow.

6.6. The resulting, prominent first reflection is intended to provide audible cues for detection of changes. All changes only affected the ERs filter, while the DD is one static filter calculated at the matched position of source and receiver. The maximum displacements for each room can be found in Table 6.5 and is limited by the room geometry for the translational displacements.

Room	Translation		Rotation
	Front	Left	
Small	6.8 m	4 m	-90°
Medium	22 m	5 m	-90°
Large	25 m	17.8 m	-90°

Table 6.5: Maximum displacements in the different rooms.

The source was simulated with a human speaker directivity [KJ99]. During the listening test participants were placed in the center of the loudspeaker array equipped with a near-field microphone (4066-OC-A-F00-LH, DPA, Allerød, Denmark) which was placed 2.5 cm away from the corner of the mouth according to the manufacturer's manual.

The task was to speak into the virtual room by using the words "one", for its open starting vocal, "two" for its impulsive, transient start and "six" for the sharp leading 's'. Participants could then freely switch between the three RIRs of the virtual rooms (of which two were the same) to identify which was one different. The RIRs were pre-calculated with a resolution of 1° for rotation and 5 cm for translation. For each scenario, the test started with the maximum displacement of the receiver. To avoid unnecessary fatigue of the participants, they were asked whether any difference is audible or not at the maximum displacement. If no difference was audible, the scenario was skipped. The test instruction emphasized to take an extra effort and time for this initial question. Additionally, a termination criterion was defined after six times of consecutive wrong identification of the different stimulus at maximum displacement. To decrease latency, 6.5 ms of leading zeros before the ground reflection at a height of 1.34 meters above ground were cut off, and the receiver was always positioned at least 1.4 meters away from the nearest wall. The GUI of the listening test was provided over a web interface running on a surface tablet and created using the PsychoPy toolbox [Pei+19]. The listening test ended with an informal, oral interview.

Results

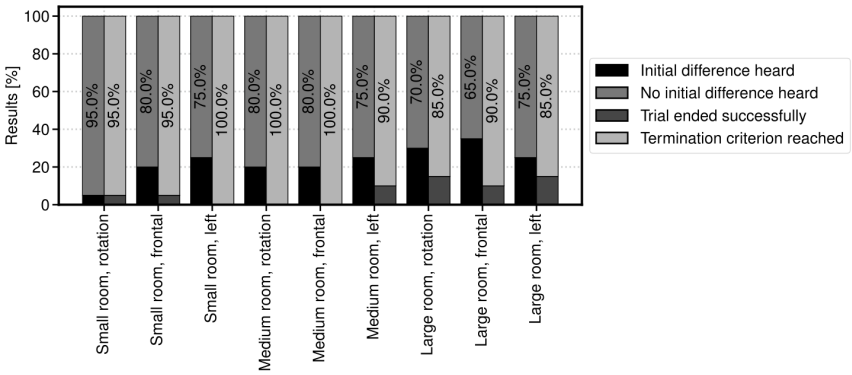


Figure 6.7: Results of the listening test. For the left bar of each condition black indicates the percentage of participants that indicated an audible difference between the reference position and the maximum mismatch and gray the ones that indicated no audible difference. For the right bar dark gray indicates that a threshold was determined (seven reversals), light gray indicates that the termination criterion applied.

20 participants with a mean age of 26.9 ($\sigma = 4.6$) years were tested in the experiment. Figure 6.7 shows the results of the listening test with an influence of 5 % per participant. For left bar of each condition the black color shows the percentage of participants that indicated that a difference between the reference and the maximum mismatch was audible. The dark gray color on the right bar indicates the percentage of participants which trials ended in a valid threshold estimation defined by a minimum number of reversals needed. Consequently, the value indicated in dark gray on the right bar can never be higher than those in black on the left. Light gray on the left bar shows the percentage of participants that indicated no audible difference for a maximum mismatch. To this number light gray on the right bar adds the participant that indicated an audible difference for the maximum mismatch but did not converge to a threshold because they run into the termination criterion. The number of participants that successfully ended a trial is too small for estimation of a mean value. Additionally, those participants with seven reversals did not necessarily converge towards one threshold but rather back to the maximum and from there one step down and up again. Single, valid thresholds are:

- small room, frontal: one participant at about 1.5 meters
- medium room, left: one participant at about 3.3 meters

- large room, rotation: two participants at about 25 degrees
- large room, frontal: one participant at about 9.5 meters

In the oral feedback interview, participants stated that the listening test was difficult as differences were rarely audible, and it was not clear whether the differences stemmed from the differences in the virtual rooms or the differences generated by different pronunciation of the same word. Last, the constant speaking was reported to be exhausting.

Discussion

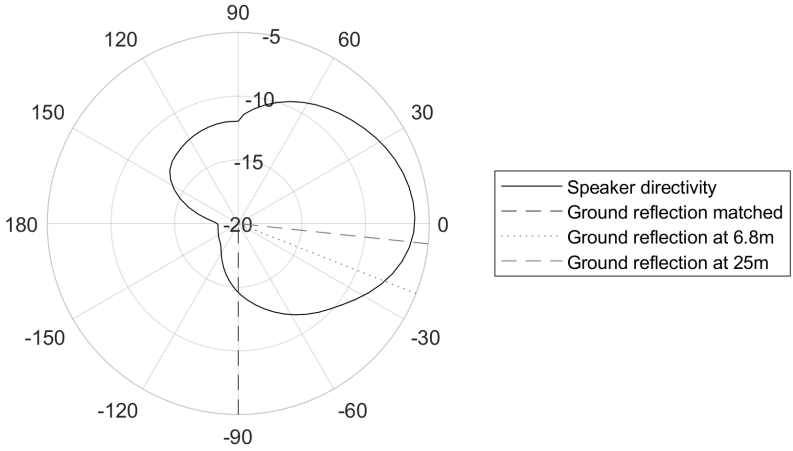


Figure 6.8: Speaker directivity in the median plane as used for the source at 1 kHz. Dotted lines indicate the angle for the ground reflection for different displacements of the receiver. The attenuation decreases from about -15 dB to around -6 dB and is contrary to the increased attenuation due to distance losses.

For the selected scenarios and the use of the speaker directivity the mismatch was not audible for most participants. While in initial tests without the speaker directivity the ground reflection delivered a main cue in terms of loudness differences, the speaker directivity compensates the distance attenuation of the ground reflection. The speaker directivity increases attenuation for directions with increasing (absolute) elevation angles (i.e., towards the top and bottom) while attenuation is low for sources closer to the horizontal plane (see Figure

6.8). The absolute elevation angle of the incoming ground reflection on the other hand decreases with increasing mismatch. Therefore, it compensates for a main part of the energy loss from the increasing length of the traveling path. By using one's own voice, participants were unable to ensure that the input signal had the same energy, which further decreases the ability to detect small loudness changes. In addition to the loudness change, the direction of the incoming reflection also changes. The ground reflection in the matched position is strongly attenuated and directed downwards to a direction without a distinct loudspeaker. For a displacement of the receiver by 6.8 m the angle rises to -21.5° in elevation, which is still close to the bottom loudspeaker ring. For a displacement of 25 meter the ground reflection arrives at an elevation angle of -6.1° which is closer to the horizontal loudspeaker ring, yet, this case only occurred for the ground reflection arriving from the rear (frontal displacement in the large room).

Therefore, in listening tests regarding self perception the update rate of the early reflections is less important as well as a temporal correct transition into the late reverberant part of the RIR. The latter effect is plausible as the early reflections shift into the diffuse part of the RIR. As the scenarios as such are missing the direct sound component and the amplitude change of the ground reflection is suppressed by a singer's directivity the differences become minimal. For the sake of completeness, it should be mentioned that even though the VR-Lab is acoustically treated, the first reflections arriving at the participant's ears are always generated from the listening room, even though with low energies. The same is true for each single virtual reflection.

Conclusion

The setup of the listening experiment was used to measure the latency occurring when interacting with the system. Assuming the first reflection is always further away than the participant's ear height, the leading zeros can be cut, and the system has a latency of 14.3 ms in a standalone mode as shown in Section 6.1.2. Participants stated no audible artifacts in the oral feedback, indicating that the system is running with sufficient computational power. Therefore, the self-perception of the user in the system can be provided by the system, allowing integration of the user into the system. The results of the listening experiment further revealed that, especially in situations where no direct comparison of the same input signal is done, the detailed time structure of ERs and DD is less important and the requirements towards the hybrid system are rather low. In the same manner, the required update rates for the ERs can be lowered and independent of those for the DD as the transition is not audible. These findings

rely mainly on the self-perception of one's own voice due to no presentation of the DS and the use of a human speaker directivity which compensates loudness changes due to the ground reflection. To further increase the test sensitivity for this scenario, a free-field environment with a single reflection in the horizontal plane (e.g., a wall) can be used. Decreasing the distance to the object of reflection would also demand a shorter delay time of the framework. For audible effects, the requirements towards the hybrid system are then higher.

7

Conclusion

During this thesis a complex construct of room acoustical simulation, manipulating of these simulation results for compensation of the listening room and, finally, combining different reproduction techniques for an overall optimized sound presentation was successfully developed that provides better localization accuracy than a pure CTC reproduction while taking benefits from VBAP and HOA in providing a better sense of naturalness, presence, speech intelligibility and, scene dependent, clarity and degree-of-linking. The system is further capable of integrating the user into the virtual scene by providing reflections of one's own voice in real-time. References to the documentation of the hybrid CVH system can be found in A. Answering the research questions from Section 1.2 this chapter concludes the insight gathered on the way: the comparison of the pure reproduction methods, the approach, and effect size of the room compensation measures and the final evaluation of the hybrid system and its capability to integrate the listener into the virtual scene. Last, a concluding suggestion towards further development is highlighted.

Findings on performance of the different reproduction systems are usually based on different loudspeaker set-ups, different implementations and different environmental conditions (especially room acoustics). The test design differs between investigations, and the vocabulary used changes. The listening experiments allow for a direct comparison between the reproduction methods in terms of loudness, provided localization accuracy and perceptual qualities. For the latter, the SAQI test method was used to provide results in a standardized format that will make results comparable with future studies, that use the same design.

The results of the loudness experiment show that variations are in a considerable range when compared to a JND of $\pm 1dB$ with single exceptions. For the frontal position, all reproduction systems differ significantly from each other. Each system shows a tendency towards the highest perception of loudness of the virtual source to the direct right side, even though only statistically significant for the HOA system. The highest variations were found for the CTC sound presentation and matches the participants' feedback, that loudness comparison is difficult if the stimuli vary in coloration. For the localization experiment, the perceived source positions varied quite high and for the elevation angle spread over the whole valid range. The 50th percentile, however, is comparable with values found in literature. The results gathered for the HOA system had to be discarded due to a technical misconfiguration. However, CTC and VBAP perform equally good in the horizontal plane where, yet, in elevation the CTC provides significantly better results for the sources to the direct right and rear right which are both in

the horizontal plane. Again, coloration was stated as an issue when comparing the direction of two different source localizations.

The comprehensive SAQI listening experiment tested three different rooms with two different scenarios each. Almost no difference between VBAP and HOA were stated by participants. The similarity of these two reproduction methods can also be found in the differences to the CTC system, which are the same for both systems. The improvements in perceiving less pronounced comb filter effects and metallic tone color can be expected (see Section 2.6.1). But both systems, VBAP and HOA, also show improved perception in the categories clarity, naturalness, presence, speech intelligibility and general degree-of-liking. Coloration might be one reason for a decreased perception of the qualities when using the CTC system. Another might be the fact, that a CTC system, because of its small sweet spot, has to continuously adapt to the participant's position and orientation on the rendering and reproduction side. As the filters for DS, ERs and DD are calculated separately, a mismatch between DS and ERs can result in the perception of ghost sources, decreasing the perceived naturalness and presence and possible further qualities. The SAQI test did not confirm the assumption that HOA (or VBAP) provide a better sense of envelopment by sound.

A novel approach to compensate ERs was presented. The listening room is modeled and simulated to compare unwanted reflections of the listening room due to loudspeaker playback with the desired reflections of the virtual room. If these reflections are close in terms of time, energy, and direction of arrival, the wanted reflection is deleted as it is already present. The approach was tested in a headphone-based listening experiment and proved effective for decreasing the perceived apparent source width. For adapting the energy decay of the overall RIR, which is the virtual room convolved with the listening room, an existing approach is used. This approach iteratively and uniformly adapts the absorption materials in the virtual room, so that the temporal structure of the RIR remains the same. However, the final evaluation did not reveal audible effects of these approaches. Partly due to the already short reverberation time of the listening room, and partly because the stimuli used might have been too long for direct comparison of the energy decay.

After the investigation of the pure reproduction systems, the novel hybrid system was implemented using CTC for the DS, VBAP for the ERs and HOA for the DD, hence, the name for the hybrid system is CVH. The results of SAQI test comparing the CVH system against the CTC system exceeded the formulated

expectations. The localization was not only maintained, but further improved. Moreover, the perceptual qualities "Naturalness", "Presence" and a higher "Speech intelligibility" and a partly improved "Clarity" and "Degree of liking" can be observed.

The hybrid system is further capable of integrating the user into the virtual scene by presenting reflections in real-time. Additionally, the acoustical integration of the user into the virtual scene can be realized with low computational requirements. The missing DS and the speaker directivity of the source (i.e., the user in the virtual scene) allow for large displacement of the user without the need for a filter update by the simulation software.

As further evaluation, the effects of moving sources on the reproduction systems should be examined. The CTC underlies constant changes of relative source to listener positions and, likely, results will not change significantly due to source movement. For VBAP and HOA on the other hand, moving sources have to be separated into two cases: pre-defined source movement can be pre-rendered entirely and may have less influence on the results. If sources have to be calculated in real-time, e.g., due to interaction with the user, the systems depend on filter update rates and the results in this thesis may change. Further, a more general suggestion for further proceedings can be found in Section 8.2.

8

Summary and outlook

8.1 Summary

A new hybrid system was introduced on the idea of separating a virtual RIR into three different sections, namely DS, ERs and DD. For each time section, a suitable reproduction method has to be selected. The initial idea is based on findings by Guastavino et al. [Gua+07] that indicate, that the reproduction systems' strengths and weaknesses are contrary. Each of them has specific strengths for the perceptual relevant qualities in one time section, but weaknesses that are only relevant for another time section of the RIR. All in all, the reproduction system should be superposed so that their strengths add up, but weaknesses are canceled (or excluded).

Three, in the scientific world established, reproduction methods were implemented and investigated: CTC, VBAP and HOA. The main criteria for using these reproduction methods is their ability to provide a full three-dimensional sound impression within a reasonable number of requirements towards the hardware, especially the number of loudspeakers needed. For HOA the rather low truncation order of two was used to provide a somewhat more enveloping sound field by a wider panning lobe, resulting in a more active loudspeaker with more homogeneous energy distribution. Additionally, the lower truncation order decreases the number of loudspeakers required. To allow ease of access for participants, a focus towards an interactive system was set, that allows free head movement and, hence, requires a dynamic binaural synthesis for the CTC reproduction method.

The first set of listening experiments were done using the pure reproduction techniques. First, a listening test to assess the loudness for each of the different reproduction systems was conducted to allow a smooth transition between the reproduction methods possible and preserve the original structure of the RIR in-between the reproduction systems. Findings in literature are difficult to compare as different investigations use different loudspeaker arrays, in different rooms, with different stimuli using different testing methods including differing perceptual qualities with different vocabulary. Furthermore, especially CTC and HOA can vary in their specific implementation as different approaches exist to realize the techniques. Therefore, listening tests were performed to achieve a direct comparison between the specific implemented reproductions systems and reveal their strengths and weakness for specific perceptual qualities. The findings provide a basis to select an appropriate reproduction system for each time slot of the RIR.

The loudness experiment compared the virtual sources to a real, reference source. Results of this experiment indicate a position dependent perception of loudness

within each reproduction system which is within the range of the JND of ±1 dB with a single exception. The CTC system results in the highest variation, and participants' feedback points towards an influence of coloration. In the localization experiment, a virtual source had to be panned into the direction of a reference loudspeaker and participants were able to switch between sound presentation of the reference source and the reproduction system. The resulting data gathered for the HOA system had to be discarded due to a technical misconfiguration. Differences between CTC and VBAP were comparable in the horizontal plane, however, the mean results in the median plane showed values close to zero for the CTC reproduction. The perceptual qualities were assessed in a direct comparison of the reproduction systems using the SAQI design. Results indicate a very similar performance of VBAP and HOA and, consequently, very similar differences towards the CTC system. The CTC system in general revealed more comb filter coloration, more metallic tone color and less clarity, naturalness, presence, speech intelligibility and general degree-of-liking. In contrast to the assumptions made towards a better sound envelopment by the listener by reverberation, the results do not show conclusive differences. Instead, they indicate a more experience of reverberation for close sources during CTC reproduction and a more distant perception of the sources.

To compensate for the listening room, two approaches, separated into one for the ERs and one more for the DD, were developed. For compensating the ERs the consideration was made, that each reflection of the simulated RIR is played back over the loudspeaker in the listening room. As the listening room itself results in a reflection of this playback, it is analyzed whether these occurring reflections can be used to replace a reflection of the virtual RIR that appears later in time. To define the suitability of these reflections, a distance metric was proposed that compares the reflection in time, intensity, and angle of density. A headphone-based, initial listening test proved that audible differences in decreasing the ASW can be created. To further optimize the compensation, the energy decay of the DD was adapted by iteratively changing the all absorption materials in the virtual room with the same factor. This approach ensures that the reflections still arrive from the same directions, with the same timing and the same relative intensity levels. The overall energy decay of the RIR of the virtual room convolved with that of the listening room then has to be matched with that of the unchanged, virtual room. Both approaches were integrated into the final hybrid system and evaluated in the subjective evaluation of the hybrid system.

The final system was selected based on the findings of the first set of listening experiments. The CTC system was chosen for more precise presentation of sound source location in the median plane. Due to the results of the prior SAQI test, the expectations towards the final system were changed. The overall better performance of perceptual qualities for VBAP and HOA reproduction are expected to improve the perception while maintaining the localization accuracy of the CTC system. Both systems, VBAP and HOA, show a very similar performance. Weak tendencies for the VBAP system to have a better performance in specific scenes and qualities were used to select this reproduction method for the ERs. For the DD the HOA system was selected to investigate whether a complete combination of all three temporal sections of the RIR is feasible and to analyze possible effects of the combining process. The final system was named CVH using the first letter of each reproduction method in chronological order.

A test scenario was developed to test different factors: the self-perception in the system as well as interaction in the system. The needed update rate for ERs due to listener movement as well as temporal accuracy needed for the transition between ERs and DD and, lastly, the setup was used to measure the framework latency. The framework latency is basis latency without any filter update latency. The measurement was done from microphone input to sound arriving at listener position and measured to 20.6 ms, including 6.7 ms for the acoustic path from loudspeaker to listener. For the listening test, the participants were equipped with a microphone and able to speak into a virtual room. The receiver position was then systematically shifted to find a threshold where the difference was audible. The results indicate that the test design allows for maximum displacement of the receiver without any noticeable difference for most participants. This may be mainly due to the test design, especially using one's own voice and not allowing for a direct A/B comparison.

The final subjective evaluation of the system used the same SAQI test as in the first set of listening tests, but reduced to each scenario with the highest deviations between the systems. It was enhanced by a localization task which compared the provided source localization against a visual reference source. Furthermore, the test included a direct comparison of the hybrid system with and without room compensation. The results of the experiment exceeded the formulated expectations towards the hybrid system: improvements in the perceptual qualities

of naturalness, presence, speech intelligibility, comb filter effects, metallic tone color and partly in degree-of-liking and clarity were found. Furthermore, source localization was further improved, instead of just as good as the CTC reproduction. The room compensation, on the other hand, proved to be not significantly audible, at least not for the already optimized VR-Lab.

8.2 Outlook

The work in this thesis provided a proof of concept of the complex hybrid system under certain constrains using specific assumptions. To further improve and validate the system, some suggestions and open questions are gathered in this section.

In general, the complexity of the system and number of variables that influences the findings is quite high. In contrast, a thorough listening focuses on as few independent parameters as possible. To further study the effects, the number of source positions, input signals as well as virtual rooms can be increased with the benefit of finding systematic changes and identify correlation or even better: causalities. An additional, very general, remark is, that the specific implementations of the systems can be changed. Different smoothing options exist for the CTC reproduction, multiple-direction amplitude panning as extensions of VBAP and various HOA decoders. Latency can be reduced by shortening the HRTF for the CTC filter calculation and defining a smaller listening area for the CTC reproduction. The filter update rates of the room acoustical simulation can be improved by using updated hardware and a separate processing unit just for the simulation.

To evaluate the effectiveness of the room compensation methods, listening rooms with higher reverberation time should be considered to see at which room size (or reverberation time), if any, the compensation becomes effective. Additionally, the weighting factors defining the distance between a reflection in the listening room and the one of the virtual RIR have to be further determined.

The reproduction methods for the ERs and the DD were chosen more or less arbitrary. It is therefore likely, that the complexity of the system can be reduced by using just one of the two. The system itself should further be tested with moving sources to see whether the delayed filter updates have the same effect on VBAP and HOA as they have on the CTC system.

Acknowledgements

I would like to express my heartfelt gratitude to everyone who supported me during the preparation of this dissertation and encouraged me to complete it.

First and foremost, I am deeply grateful to my supervisor, Professor Michael Vorländer, for his invaluable guidance, insightful feedback, and unwavering support throughout this journey. His expertise and constructive suggestions were instrumental in shaping this dissertation, and his leadership profoundly influenced my personal and professional growth.

I extend my sincere gratitude to Professor Bruno Sanches Masiero for serving as my second supervisor, carefully reviewing my thesis, and providing valuable feedback that enriched my work.

Special thanks to Gottfried Behler for always being available with valuable advice, offering a calm perspective that had a lasting impact on both this dissertation and me personally.

I am also grateful to my colleagues at the institute for their collegiality, stimulating discussions, and constant support - especially during listening tests amid COVID - 19—which made this experience both enriching and memorable.

My sincere appreciation goes to the Acoustic Virtual Reality group—Lukas Aspöck, Sönke Pelzer, Jonas Stienen, and Frank Wefers—for their professional and personal support. A heartfelt thank you to Marco Berzborn, Ramona Bomhardt, Johannes Klein, and Rob Opdam for the climbing, running, and the balance and perspective you bring to my life.

Additionally, I deeply value the engaging discussions, both inside and outside the institute, with Elie Abi-Raad, Hark Braren, Simon Kersten, Mark Müller-Giebeler, Josefa Oberem, Jan Richter, Pascal Palenda, Florian Pausch, Philipp Schäfer, Ingo Witew and Manuj Yadav. Your insights and camaraderie have been truly invaluable.

I am sincerely thankful to the secretariat team - Ellen Vergöls, Kim König, Emilie Koch - and especially Karin Charlier, for their invaluable administrative support. My gratitude also extends to the electrical workshop - Rolf Kaldenbach, Norbert Konkol - and the mechanical workshop - Marc Eiker, Thomas Schaefer, Uwe Schlömer - for their expertise and practical solutions.

I appreciate the dedication of the student assistants and thesis contributors, particularly Erik Röcher and Markus Voth. Working with Nils Rummler was a privilege; his independence, reliability, and expertise in professional audio were invaluable. Thank you for embracing my experimental ideas with patience.

Above all, I want to express my deepest gratitude to my family. To my wife, Wiebke, for her endless patience, encouragement, and for always having my back, even during the most challenging times. To my children, Lina and Nele, who have been a constant source of joy and motivation throughout this journey. And to my parents, Hannelore and Hans-Dieter Kohnen, for their unconditional love and the values of perseverance and dedication they installed in me, which have been my foundation.

To everyone who contributed to this journey - whether through academic guidance, critical feedback, or personal support - thank you. This dissertation is not only the result of my efforts, but also a reflection of the inspiring and supportive environment I was fortunate to be part of.

Curriculum Vitae

Nov. 5th, 2024	Doctoral Examination for Dr.-Ing.
2015 - 2023	Research Assistant and Doctoral Candidate Institute for Hearing Technology and Acoustics RWTH Aachen University, Germany
2014 - 2015	Graduate Assistant Institute for Hearing Technology and Acoustics RWTH Aachen University, Germany
March, 2014	Master of Science, Electrical Engineering, Information Technology and Computer Engineering RWTH Aachen University, Germany
March, 2012	Bachelor of Science, Electrical Engineering, Information Technology and Computer Engineering RWTH Aachen University, Germany
July, 2006	General Higher Education Entrance Qualification (Abitur) St. Georg Gymnasium, Bocholt, Germany
1986	born in Bocholt, Germany

A

Appendix

Hybrid system

The documentation of the hybrid system in the most recent version can be found here:
https://zenodo.org/doi/10.5281/zenodo.7128184

The documentation at the current state (v0, Oct 12th, 2022) of the thesis:
https://doi.org/10.5281/zenodo.7128185

Statistical data for listening experiments

The SPSS syntax and output of for the listening tests in their most recent version can be found here:
https://zenodo.org/doi/10.5281/zenodo.12740450

The data at the current state of the thesis (v0, Jul 14th, 2024):
https://doi.org/10.5281/zenodo.12740451

Bibliography

[AB79] J. B. Allen and D. A. Berkley. “Image method for efficiently simulating small-room acoustics”. In: *Journal of the Acoustical Society of America* 65.4 (1979), pp. 943–950. ISSN: NA. DOI: 10.1121/1.382599.

[Asp+20] L. Aspöck, F. Brinkmann, D. Ackermann, S. Weinzierl, and M. Vorländer. *BRAS - Benchmark for Room Acoustical Simulation, additionalData*. 2020. DOI: https://doi.org/10.14279/depositonce-6726.3.

[Bar09] M. Barron. *Auditorium Acoustics and Architectural Design*. 2nd ed. Spon Press, 2009. ISBN: 9781135219260. DOI: 10.4324/9780203874226.

[BWA01] D. R. Begault, E. M. Wenzel, and M. R. 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: *108th, Audio Engineering Society Convention*. Vol. 49. 10. Paris, France, 2001, pp. 904–916.

[Ber+17] M. Berzborn, R. Bomhardt, J. Klein, J.-G. Richter, and M. Vorländer. “The ITA-Toolbox: An open source MATLAB toolbox for acoustic measurements and signal processing”. In: *Proceedings of the 43th Annual German Congress on Acoustics, DAGA 2017*. Kiel, Germany, 2017, pp. 222–225. ISBN: 978-3-939296-12-6.

[Bla96] J. Blauert. *Spatial Hearing*. Revised. Cambridge, Massachusetts: The MIT Press, 1996. ISBN: 9780262268684. DOI: 10.7551/mitpress/6391.001.0001.

[Bla13] J. Blauert, ed. *The technology of binaural listening*. Modern acoustics and signal processing. Heidelberg et al.: Springer, 2013.

[Bri+19] F. Brinkmann, L. Aspöck, D. Ackermann, S. Lepa, M. Vorländer, and S. Weinzierl. “A round robin on room acoustical simulation and auralization”. In: *The Journal of the Acoustical Society of America* 145.4 (2019), pp. 2746–2760. DOI: 10.1121/1.5096178.

[Bri+21] F. Brinkmann, L. Aspöck, D. Ackermann, R. Opdam, M. Vorländer, and S. Weinzierl. "A benchmark for room acoustical simulation. Concept and database". In: *Applied Acoustics* 176 (2021), p. 107867. ISSN: 0003682X. DOI: 10.1016/j.apacoust.2020.107867.

[Car17] T. Carpentier. "Normalization Schemes in Ambisonic: Does it Matter?" In: *Journal of The Audio Engineering Society*. 2017.

[Cre77] L. Cremer. "Law of the First Wave Front". In: *Journal of the Audio Engineering Society* 25.6 (1977), pp. 420, 422.

[Dan00] J. Daniel. "Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia". PhD thesis. Université Pierre et Marie Curie (Paris VI), 2000.

[DRP98] J. Daniel, J.-B. Rault, and J.-D. Polack. "Ambisonics Encoding of Other Audio Formats for Multiple Listening Conditions". In: *105th Audio Engineering Society Convention*. 1998.

[Emm+88] D. S. Emmerich, J. Harris, W. S. Brown, and S. P. Springer. "The relationship between auditory sensitivity and ear asymmetry on a dichotic listening task". In: *Neuropsychologia* 26.1 (1988), pp. 133–143. ISSN: 00283932. DOI: 10.1016/0028-3932(88)90036-X.

[FB10] S. Favrot and J. M. Buchholz. "LoRA: A Loudspeaker-Based Room Auralization System". In: *Acta Acustica united with Acustica* 96.2 (2010), pp. 364–375. ISSN: 1610-1928. DOI: 10.3813/AAA.918285.

[Fra14] M. Frank. "How to make Ambisonics sound good". In: *Forum Acusticum*. European Acoustic Association. Kraków, Poland, 2014.

[GHG24] M. Gerken, V. Hohmann, and G. Grimm. "Comparison of 2D and 3D multichannel audio rendering methods for hearing research applications using technical and perceptual measures". In: *Acta Acustica* 8 (2024), p. 17. ISSN: 2681-4617. DOI: 10.1051/aacus/2024009.

[GP15] J. Grosse and S. van de Par. "Perceptually Accurate Reproduction of Recorded Sound Fields in a Reverberant Room Using Spatially Distributed Loudspeakers". In: *IEEE Journal of Selected Topics in Signal Processing* 9.5 (2015), pp. 867–880. ISSN: 1932-4553. DOI: 10.1109/JSTSP.2015.2402631.

[Gua+07] C. Guastavino, V. Larcher, G. Catusseau, and P. Boussard. "Spatial Audio Quality Evaluation: Comparing Transaural, Ambisonics and Stereo". In: *Proceedings of the 13th International Conference on Auditory Display* (2007), pp. 53–58. ISSN: 1520-8524.

[Ins] Institute for Hearing Technology and Acoustics. *Virtual Acoustics - A real-time auralization framework for scientific research*. URL: http://www.virtualacoustics.org/ (visited on 07/18/2024).

[Kin+94] P. E. King-Smith, S. S. Grigsby, A. J. Vingrys, S. C. Benes, and A. Supowit. "Efficient and unbiased modifications of the QUEST threshold method:

theory, simulations, experimental evaluation and practical implementation". In: *Vision Research* 34.7 (1994), pp. 885–912. ISSN: 0042-6989.

[KJ99] M. Kob and H. Jers. "Directivity Measurement of a Singer". In: *Journal of the Acoustic Society of America* 105.2 (1999), p. 1003.

[Koh+18] M. Kohnen, R. Bomhardt, J. Fels, and M. Vorländer. "Just noticeable notch smoothing of head-related transfer functions". In: *Proceedings of the 44th Annual German Congress on Acoustics, DAGA 2018*. Munich, Germany, 2018, pp. 333–335.

[Koh+21] M. Kohnen, F. Denk, J. Llorca-Bofi, B. Kollmeier, and M. Vorländer. "Cross-site investigation on head-related and headphone transfer functions: variabilities in relation to loudness balancing". In: *Acta Acustica* 5 (2021), p. 58. ISSN: 2681-4617. DOI: 10.1051/aacus/2021051.

[Koh+16] M. Kohnen, J. Stienen, L. Aspöck, and M. Vorländer. "Performance Evaluation of a Dynamic Crosstalk-Cancellation System with Compensation of Early Reflections". In: *2016 AES International Conference on Sound Field Control*. 2016.

[KSV17] M. Kohnen, J. Stienen, and M. Vorländer. "Subjective evaluation of a room-compensated crosstalk cancellation system". In: *Proceedings of the 43th Annual German Congress on Acoustics, DAGA 2017*. Kiel, Germany, 2017.

[KV19a] M. Kohnen and M. Vorländer. "Loudness differences between different reproduction techniques". In: *Proceedings of the 45th Annual German Congress on Acoustics, DAGA 2019*. Rostock, Germany, 2019, pp. 535–538.

[KV19b] M. Kohnen and M. Vorländer. "Virtual scene adaption for compensation of the reproduction room". In: *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*. Madrid, Spain: Institute of Noise Control Engineering, 2019, pp. 3592–3601.

[KV21] M. Kohnen and M. Vorländer. "Perceptional differences between Transaural, Ambisonics and Amplitude Panning". In: *Proceedings of the 47th Annual German Congress on Acoustics, DAGA 2021*. Wien, Austria, 2021.

[Kut16] H. Kuttruff. *Room Acoustics*. 6th ed. Boca Raton: CRC Press, 2016. ISBN: 9781315372150. DOI: 10.1201/9781315372150.

[Len06] T. Lentz. "Dynamic Crosstalk Cancellation for Binaural Synthesis in Virtual Reality Environments". In: *Journal of The Audio Engineering Society* 54.4 (2006), pp. 283–294.

[Lin14] A. Lindau. *Spatial Audio Quality Inventory (SAQI). Test Manual*. Online. 2014. URL: http://dx.doi.org/10.14279/depositonce-1.2.

[Lin+14] A. Lindau, V. Erbes, S. Lepa, H.-J. Maempel, F. Brinkmann, and S. Weinzierl. "A Spatial Audio Quality Inventory (SAQI)". In: *Acta Acustica united with Acustica* 100.5 (2014), pp. 984–994.

[LKW12] A. Lindau, L. Kosanke, and S. Weinzierl. “Perceptual evaluation of model- and signal-based predictors of the mixing time in binaural room impulse responses”. In: *Journal of the Audio Engineering Society* 60.11 (2012), pp. 887–898.

[LGF05] J. Lopez, A. Gonzalez, and L. Fuster. “Room compensation in wave field synthesis by means of multichannel inversion”. In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005.* IEEE, 2005, pp. 146–149. ISBN: 0-7803-9154-3. DOI: 10.1109/ASPAA.2005.1540190.

[MMF13] P. Majdak, B. Masiero, and J. Fels. “Sound localization in individualized and non-individualized crosstalk cancellation systems”. In: *Journal of the Acoustic Society of America* 133.4 (2013), pp. 2055–2068.

[Mas12] B. Masiero. “Individualized binaural technology. Measurement, equalization and perceptual evaluation”. Doctoral Thesis. RWTH Aachen University, Germany, 2012.

[Møl+96] H. Møller, M. F. Sørensen, C. B. Jensen, and D. Hammershøi. “Binaural Technique: Do We Need Individual Recordings?” In: *Journal of the Audio Engineering Society* 44.6 (1996), pp. 451–469.

[PKV23] P. Palenda, M. Kohnen, and M. Vorländer. “Influence of Position Mismatch on the Perception of Early Reflections of One’s Own Voice”. In: *Proceedings of the 49th Annual German Congress on Acoustics, DAGA 2023*. Hamburg, Germany, 2023.

[Pau22] F. Pausch. *Documentation of the experimental environments and hardware used in the dissertation "Spatial audio reproduction for hearing aid research: System design, evaluation and application"*. Tech. rep. Aachen, Germany, 2022. DOI: 10.18154/RWTH-2022-01536.

[Pei+19] J. Peirce, J. R. Gray, S. Simpson, M. MacAskill, R. Höchenberger, H. Sogo, E. Kastman, and J. K. Lindeløv. “PsychoPy2: Experiments in behavior made easy”. In: *Behavior Research Methods* 51.1 (2019), pp. 195–203. ISSN: 15543528. DOI: 10.3758/s13428-018-01193-y.

[PV13] S. Pelzer and M. Vorländer. “Auralization of virtual rooms in real rooms using multichannel loudspeaker reproduction”. In: *The Journal of the Acoustical Society of America*. Vol. 134. 5. 2013, pp. 3985–3985. DOI: 10.1121/1.4830520.

[PSV11] S. Pelzer, B. Sanches Masiero, and M. Vorländer. “3D reproduction of room acoustics using a hybrid system of combined crosstalk cancellation and ambisonics playback”. In: *Proceedings of ICSA*. 2011, pp. 297–301.

[PSV14] S. Pelzer, B. Sanches Masiero, and M. Vorländer. “3D Reproduction of Room Auralizations by Combining Intensity Panning, Crosstalk Cancellation and Ambisonics”. In: *Proceedings of the EAA Joint Symposium on Auralization and Ambisonics 2014*. Berlin: Universitätsbibliothek Technische Universität Berlin, 2014, pp. 182–188. DOI: 10.14279/depositonce-33.

[Pul97] V. Pulkki. “Virtual sound source positioning using vector base amplitude panning”. In: *Journal of the Audio Engineering Society* 45.6 (1997), pp. 456–466.

[Pul01] V. Pulkki. “Localization of Amplitude-Panned Virtual Sources II: Two- and Three-Dimensional Panning”. In: *Journal of the Audio Engineering Society* 49.9 (2001), pp. 753–767.

[Ric19] J.-G. Richter. “Fast Measurement of Individual Head-Related Transfer Functions”. PhD thesis. RWTH Aachen University, Germany, 2019.

[Rie98] K. A. J. Riederer. “Repeatability analysis of head-related transfer function measurements”. In: *105th Audio Engineering Society Convention*. San Francisco, USA, 1998.

[Sch95] A. Schmitz. “Ein neues digitales Kunstkopfmeßsystem”. In: *Acta Acustica united with Acustica* 81.4 (1995), pp. 416–420.

[Sch11] D. Schröder. “Physically based real-time auralization of interactive virtual environments”. PhD thesis. RWTH Aachen University, Germany, 2011.

[SV11] D. Schröder and M. Vorländer. “RAVEN: A real-time framework for the auralization of interactive virtual environments”. In: *Forum Acusticum*. Aalborg Denmark. 2011, pp. 1541–1546.

[SBR06] S. Spors, H. Buchner, and R. Rabenstein. “Eigenspace adaptive filtering for efficient pre-equalization of acoustic MIMO systems”. In: *2006 14th European Signal Processing Conference*. Florence, Italy, 2006.

[TZA14] D. S. Talagala, W. Zhang, and T. D. Abhayapala. “Efficient Multi-Channel Adaptive Room Compensation for Spatial Soundfield Reproduction Using a Modal Decomposition”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22.10 (2014), pp. 1522–1532. ISSN: 2329-9290. DOI: 10.1109/TASLP.2014.2339195.

[TAK17] L. Thresh, C. Armstrong, and G. Kearney. “A Direct Comparison of Localisation Performance When Using First, Third and Fifth Order Ambisonics For Real Loudspeaker and Virtual Loudspeaker Rendering”. In: *Journal of The Audio Engineering Society*. 2017.

[Vor20] M. Vorländer. *Auralization*. 2nd ed. Cham: Springer International Publishing, 2020. ISBN: 978-3-030-51201-9. DOI: 10.1007/978-3-030-51202-6.

[WP83] A. B. Watson and D. G. Pelli. “Quest: A Bayesian adaptive psychometric method”. In: *Perception and Psychophysics* 33.2 (1983), pp. 113–120. ISSN: 0031-5117. DOI: 10.3758/BF03202828.

[ZF19] F. Zotter and M. Frank. *Ambisonics*. Vol. 19. Springer Topics in Signal Processing. Cham: Springer International Publishing, 2019. ISBN: 978-3-030-17206-0. DOI: 10.1007/978-3-030-17207-7.