Master's Thesis:

# Listening experiment on the plausibility of acoustic modeling in virtual reality

**Kajetan Simon Enge, BSc**
Matr.Nr.: 01230641
Supervisor: DI Ph.D. Matthias Frank
Assessor: O.Univ.Prof. Mag.art. DI Dr.techn. Robert Höldrich
University of Music and Performing Arts Graz
University of Technology Graz
Master's Degree Programme: Electrical Engineering and Audio Engineering
Graz, April 19, 2020


# Statutory declaration

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

Graz, April 19, 2020 Kajetan Simon Enge, BSc

# Kurzfassung

Virtual Reality Umgebungen werden in verschiedensten Anwendungsbereichen wie der Architektur, dem Film, in Computerspielen, in der Bildung, der Medizin und in therapeutischen Methoden zu einem immer wichtigeren Faktor. VR Anwendungen konzentrierten sich lange primär auf visuelle Eindrücke. Unbestritten ist allerdings, dass dreidimensionaler Klang eine bedeutende Rolle bei der Glaubwürdigkeit virtueller Umgebungen spielt, alleine schon aus dem Grund, dass realer Raum vom Menschen maßgeblich über seinen Klang charakterisiert wird. Das Pendant einer VR-Brille für die Augen ist für die Ohren die binaurale Wiedergabe über Kopfhörer. Die Virtualisierung realer Räume mittels Programmen wie Unity oder der Unreal Engine eröffnet neue Möglichkeiten für psychoakustische Untersuchungen, nämlich mit binauralem Rendering in virtuellen Räumen. In der vorliegenden Arbeit wurde die wahrgenommene Plausibilität verschiedener Kombinationen aus visuellen und akustischen Virtualisierungen realer Räume untersucht. Diese Kombinationen wurden von 20 Versuchspersonen mit vier verschiedenen Graden von Bewegungsfreiheit bewertet. Es zeigt sich unter anderem, dass volldynamische akustische Modellierungen mit Ambisonics dritter Ordnung gleich gute Plausibilitätsbewertungen erzielen wie Modellierungen mit siebter Ordnung. Auch eine statische Modellierung des gesamten Nachhalls mittels Faltung einer BRIR erzielt ähnliche Ergebnisse, solange der Direktschall dynamisch präsentiert wird. Wenn der akustisch präsentierte und der visuell präsentierte Raum einen stark unterschiedlichen Größeneindruck vermitteln sinkt die Plausibilität.

# Abstract

Virtual Reality environments are becoming an increasingly important factor in various applications such as architecture, film, computer games, education, medicine, and therapeutic methods. For a long time, VR applications focused primarily on visual impressions. It is undisputed that three-dimensional sound plays an essential role in the credibility of virtual environments, if only for the reason that real space is largely characterized through its sound. The counterpart of VR glasses for the eyes is binaural reproduction via headphones for the ears. The virtualization of real spaces using programs such as Unity or the Unreal Engine opens up new possibilities for psychoacoustic investigations, namely with binaural rendering in virtual spaces. In the present work, the perceived plausibility of different combinations of visual and acoustic virtualizations of several real spaces was investigated. Twenty participants evaluated these combinations with four different degrees of freedom. Among other things, it is shown that fully dynamic third-order Ambisonics models achieve the same plausibility as models with seventh order. Also, static modeling of the entire reverberation by a convolution of a BRIR reverb delivers similar results as long as the direct sound is presented dynamically. The plausibility decreases if the acoustics and the visuals provide a strongly different room size impressions.

# Contents

# Chapter 1

# Introduction

Virtual reality is a technology that has evolved very quickly in recent years and is expanding into more and more areas of application. These areas range from architecture, industry, film, gaming, tourism, education, safety training and medicine to experiments in therapeutic methods for Alzheimer's, schizophrenia or autism [MA20, BV17, NHBH20, BMJ20, KA03, JEK19]. In some of these fields, the desire for high-quality sound in VR is also increasing. Well, what do people understand by high-quality sound in the context of virtual reality? Here it is not only about audio production in the classical terms of music production. Since people in VR can perceive their visual environment via a head-mounted display (HMD) that simulates a three-dimensional space for their eyes, the sound must also be three-dimensional. This thesis examines different acoustic modeling techniques for their perceived plausibility in VR applications. This chapter first introduces the technology of virtual reality, then deals with terms such as "presence," "immersion," "plausibility," and "authenticity" and applies these terms to virtual acoustics. It then explains how this thesis contributes to the research area and introduces the basic methodology of the project.

## 1.1 Virtual Reality

Petr Kellnhofer describes in [Kel16] the technical requirements and also the associated problems for HMDs in the presentation of virtual spaces. Our brain uses the small difference in perspective between our left and right eye to perceive visual depth. Figure 1.1 shows how HMDs use this concept. HTC VIVE [1] or Oculus [2] are two companies providing such HMDs. In VR applications the virtual environments are often designed with game development engines such as Unity [3] or Unreal Engine [4]. These engines offer the possibility to connect an HMD, allowing the users to explore any virtual scene. Usually, the goal of the VR applications is to give the user the feeling as if they were there them-

1. https://www.vive.com
2. https://www.oculus.com
3. https://unity.com
4. https://www.unrealengine.com

selves. How it is possible to make people feel this way with this technology has become a subject of research in various disciplines. In this context, terms such as "spatial presence," "virtual embodiment," or "involvement" are being investigated [MBWS06,Hof16,Kas20]. Since a detailed examination of these topics in a broad context would go beyond the scope of this thesis, the terms "immersion," "presence," "plausibility," and "authenticity" are discussed here, especially with respect to acoustics.

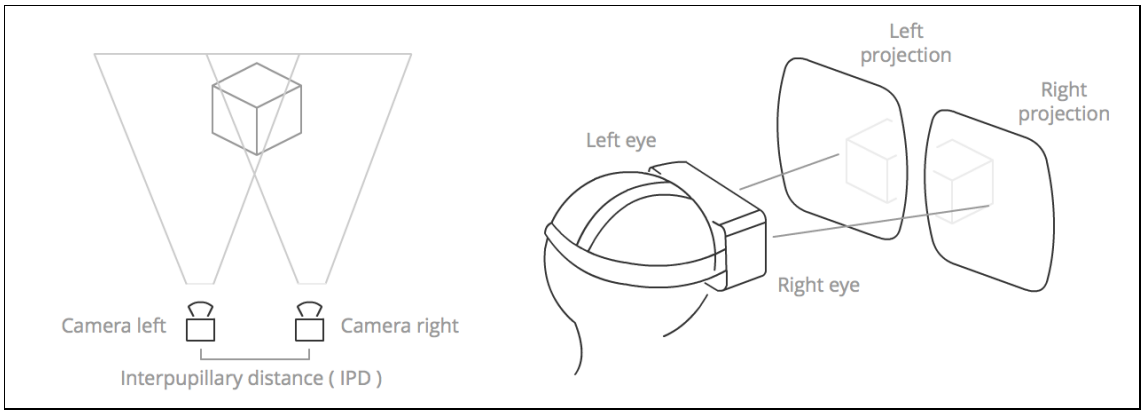



Figure 1.1 – The concept for the design of a stereoscopic image in a HMD. Source: `https://developer.mozilla.org/en-US/docs/Web/API/WebVR_API/Concepts`

Mel Slater differentiates between the two words "immersion" and "presence". He states in [Sla03, p.1]:

> *"I have argued before about the separation of the term 'immersion' from 'presence' [...]. Let's reserve the term 'immersion' to stand simply for what the technology delivers from an objective point of view. The more that a system delivers displays (in all sensory modalities) and tracking that preserves fidelity in relation to their equivalent real-world sensory modalities, the more that it is 'immersive'."*

Within this context, VR environments deliver mainly (and in most cases only) displays for visual modalities. At the current state, most VR applications ignore the other human senses. Still, the same environments try the reach what is called "presence," being "*the response to a given level of immersion*", according to Mel Slater in [Sla03, p.5]. What he says is that presence is a state the users can reach in a virtual environment when the technology provides the necessary immersion. It is their reaction to the immersive presentation. Note that humans perceive physical space to some extent through its acoustics. One only has to walk into a room with closed eyes and clap both hands to get an instant impression of the surrounding. Moreover, the sense of hearing is omnidirectional, whereas people can see only in the frontal direction. Considering this fundamental relationship between space and sound, it seems almost absurd to try to create a state like presence without the immersive potential of acoustics.

Another interesting pair of vocabulary is "plausibility" and "authenticity." Kuhn-Rahloff defines plausibility in [KR11, p. 129]:

> *"Plausibility is the result of a perceptive process that determines the extent to which an object of perception corresponds to an inner reference resulting from individual previous experiences."*

Whenever the extent of this correspondence between the perception of an object and the inner reference is maximal, the result is authenticity. Authenticity, therefore, describes a reproduction that is indistinguishable from reality [Pel01]. This creates a quite beautiful analogy: When a technical system provides maximum immersion, presence results. When a system is modeling reality with maximum plausibility, then authenticity results. The question regarding virtual acoustics requires to be: "Which parameters of real acoustics have to be virtualized to what extent in order to enable an immersion that creates presence?" In other words: "What needs to be modeled to provide virtual acoustics as plausible as an authentic acoustic scene?"

## 1.2 The contribution of this study

In the context of audio in virtual reality, the present work contributes to the understanding of the plausibility of different modeling techniques. For this purpose, a listening experiment was conducted with 20 participants, that investigated the plausibility of 68 different stimuli. Those stimuli were combined out of six different visual conditions, ten different acoustic conditions, and four different levels of degrees of freedom.

On the one hand, this study investigates the question: "How would one have to realize an acoustic model of a virtual room to provide the people in VR with a plausible audiovisual impression?" On the other hand, the setup makes it possible to present combinations of acoustic and visual impressions that do not exist in reality, i.e., a small studio room with a very long reverb or a concert hall with very short reverb. Stephan Werner et al. investigated such situations with binaural recordings of a loudspeaker in one room and the video projection of a different room as a visual condition in [WKMB16]. Their study suggests that the visual impression influences the externalization of binaural reproduction; they called it the "room divergence effect." Inspired by Werner's findings, this thesis investigates something we can call "virtual room divergence effect." This would mean that the visual impression of a room influences the plausibility of its acoustic representation.

In order to investigate these two questions, the participants rated the plausibility of the acoustic impression of a virtual and a real loudspeaker in several different visual environments. An extract from this thesis was also published in [EFH20].

## 1.3 Ambisonics and binaural audio in the context of VR

The human perception of real acoustics is influenced for example by characteristics like reverberation, localization, timbre, and dynamics. Both the sound source and its environment affect the acoustic perception in any situation. For the acoustic modeling in a VR environment it is also interesting to model both the source's and and the the room's acoustic behavior. What actually has to be taken into account to present plausible virtual acoustics?

**Modeling human spatial hearing with binaural & Ambisonics technology:**
Just like the fact that having two eyes enables the human vision to perceive spatial depth, the fact that we have two ears enables the human auditory system to perceive spatial sound. It was Rayleigh in [Ray07] who first published a theory - called Duplex theory - on how we perceive spatial sound. He describes the inter-aural time difference (ITD) to be responsible for our perception of direction for low frequencies and the inter-aural level differences (ILD) to be responsible for high frequencies. The ILD and ITD cannot provide enough information to find the location of the source in the whole three-dimensional space. A so-called "cone of confusion" describes an area from which a sound would cause the same ITDs and ILDs. The reflections from the pinna, shoulders, and torso provide additional spectral information to the listener to distinguish also between directions from one cone of confusion [CMM05]. The idea behind binaural technology is to provide spatial sound via headphones. This is possible because of our knowledge of the so-called head-related transfer function (HRTF). The HRTF is a description of the relation between the sound pressure in someone's ear canal to the sound pressure at the center of their head if the person was absent. This incorporates all the parameters from above: ITD, TLD, and spectral differences due to reflections. If we record a sound that includes these features, then playing it back via headphones will cause a three-dimensional acoustic impression. An audio signal will offer these features to the listener if it was (I) recorded with an artificial head microphone like the KU 100 [5], (II) if it was recorded with small microphones in the ears of somebody [6], or (III) if it was convolved with a binaural room impulse response (BRIR). An impulse response generally describes the relation of any system's output to its input. In this case the system is the combination out of the room acoustics and the binaural perception. Next to the fact that dynamic binaural rendering is helpful to minimize localization confusions [Ger73], it is especially crucial for VR environments because they take the head movements of the users into account. There are two basic concepts of how to realize dynamic rendering. The one is to measure a dense set of BRIRs from many spatial directions and to interpolate between them according to the head rotations of the user [ZZF19, LR]. The second technique to render dynamic binaural audio is described in [ZSH18] and generates comparable results but with far less effort. It uses the advantages of another technology that (re-)produces three-dimensional sound: Ambisonics [Ger73, ZF19].

5. https://de-de.neumann.com/ku-100
6. https://www.dpamicrophones.com/immersive/4560-core-binaural-headset-microphone

Ambisonics usually provides three-dimensional sound by encoding any sound field to its Spherical Harmonics representation and by then decoding this representation to an array of loudspeakers. Ambisonics is the perfect technology for VR acoustics because (I) an arbitrary rotation of the entire sound field in the Ambisonics domain can be done via a rotation matrix [ZF19] and (II) the signals can also be decoded binaurally for headphones [ZSH18, ZF19].

$Y_n^m$ in Equation 1.1 are the Spherical Harmonics of order n and degree m with $P_n^{|m|}$ being the associated Legendre Polynomials (see equation 1.2). These associated Legendre Polynomials describe the dependency of the Spherical Harmonics on the zenith angle. The angular terms - $\sin(|m|\varphi)$ if $m < 0$ and $\cos(|m|\varphi)$ if $m \geq 0$ - take care of the dependency on the azimuth angle. $P_n^{|m|}(\cos\vartheta)$ itself is dependent on $P_n(\cos\vartheta)$, the unassociated Legendre Polynomials (see equation 1.3). For Ambisonics, the normalization term $N_n^m$ is usually chosen to be $\sqrt{\frac{1}{2}\frac{(n-|m|)!}{(n+|m|)!}}$, which is then called a SN3D normalization [ZF19]. The unassociated Legendre Polynomials will become important later in this thesis, when a directivity pattern for the virtual loudspeaker is calculated.

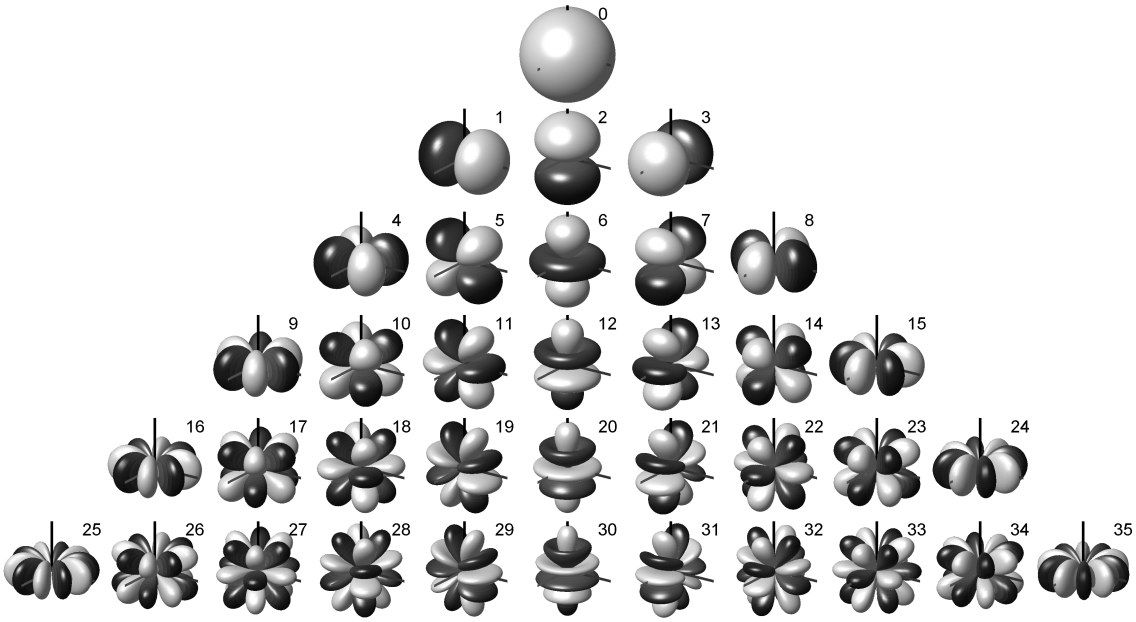



Figure 1.2 – The Spherical Harmonics up to 5th order.

$$Y_n^m = N_n^m P_n^{|m|}(\cos\vartheta) \begin{cases} \sin(|m|\varphi) & \text{if } m < 0. \\ \cos(|m|\varphi) & \text{if } m \geq 0. \end{cases} \tag{1.1}$$

$$P_n^{|m|}(\cos\vartheta) = (-1)^m (1-\cos^2\vartheta)^{m/2} \frac{d^m}{d(\cos\vartheta)^m}(P_n(\cos\vartheta)) \tag{1.2}$$

$$P_n(\cos\vartheta) = \frac{1}{2^n\, n!} \frac{d^n}{d(\cos\vartheta)^n}(\cos\vartheta^2 - 1)^n \tag{1.3}$$

**Trends in research for acoustics in virtual and augmented reality:**
Virtual acoustics is a field of research developing fast at the moment. This is partly due to the fast evolution of VR and AR (augmented reality) and the growing demand for plausible (ultimately authentic) acoustics within these applications. One current example of augmented acoustic reality is the "Augmented Practice-Room" [FRB20]. This application simulates variable room acoustics in real-time via open headphones, while the direct sound of an instrument can be heard unmodified. Zhenyu et al. use neural networks to estimate acoustic parameters from recordings of real rooms and to create virtual acoustics with the estimations [TBL+20]. Pulkki and Svensson use neural networks to find filter parameters for the modeling of scattering from geometric objects in virtual reality [PS19].

Something else that is currently being studied a lot are applications with so-called "six degrees of freedom" (6 DoF). They allow the user to freely walk through a room and listen to interpolated Ambisonics recordings that were recorded with a grid of microphones in a real room [ZFSH20]. This ultimately allows the users to walk through and listen to an acoustic scene, that was recorded earlier in a different room.

## 1.4 Basics of the visual environments

The next chapter will explain the technical setup of the listening experiment. To be able to understand the necessities of some technical implementations, it is important to know about the general differences in the visual stimuli. There were two groups: virtual reality and reality. The virtual reality environments were presented via a head-mounted display. For visual reality, the participants used their own eyes to perceive their environment. The presentation of the acoustic environments was done via open headphones and generally depended on the position and orientation of the HMD. Since the HMD cannot provide this information when the participant is not using it, it was necessary to implement a second system to gather the position and orientation data. This was an OptiTrack [7] system.

Within the experiment, five of the six visual environments were virtual. The one real environment was the studio room in which the experiment took place. In the real room, also a real loudspeaker was placed in the center of the room. In the virtual rooms, a virtual version of that same loudspeaker was shown at the same position. This means that in relation to the participant, the virtual or real loudspeaker was always in the same position. One of the virtual rooms was a representation of the real studio room. Here, not only the loudspeaker but all of the interior of the studio was virtualized. This way, the participants were able to move through the same room in reality and virtual reality. The remaining visual environments were virtualizations of different rooms at the University of Music and Performing Arts Graz.

7. https://optitrack.com

# Chapter 2

# Technical Test Setup

To be able to test a wide variety of combinations of acoustic and visual stimuli, it was necessary to implement a very flexible test system. This chapter describes the different technical systems that interacted with each other to realize the listening experiment. The following hardware and software played a role in the setup:

Hardware:

- One HTC VIVE head mounted display [1] (HMD).
- One pair of open headphones.
- Two HTC VR tracking cameras.
- Eight OptiTrack Flex 13 tracking cameras [2] and two USB Hubs.
- One OptiTrack reflection object.
- One computer for running Unity, Steam VR, Pure Data, and Reaper.
- One computer for running Motive (tracking software).

Software:

- Unity [3] (Visual rendering and test management)
- Reaper [4] (Acoustic rendering)
- Pure Data [5] (Visual interface in real room)
- Steam VR [6] (Interface between Unity and HMD)
- Motive [7] (OptiTrack)

1. https://www.vive.com/eu/product/vive/
2. https://optitrack.com
3. https://unity.com
4. http://reaper.fm
5. https://puredata.info
6. https://store.steampowered.com/app/250820/SteamVR
7. https://optitrack.com/products/motive/

One can imagine that the technical setup was a complex one. The next section explains how this system was set up and how the different units communicated with each other.

## 2.1 VR Setup in the Room

The room in which the listening experiment was conducted was one of the studio rooms of the IEM (room number PT116EG012 at Petersgasse 116, Graz). It is a room with the measures 6m $\times$ 4.4m $\times$ 3.2m (depth $\times$ width $\times$ height). Figure 2.1 shows a picture of the studio. In this figure, red circles, rectangles, and hexagons mark the essential technical setup. The one window in the room faces south-east. On the walls left and right from the window, big acoustic absorbers are mounted. These absorbers were used to attach the cameras that tracked the participants during the listening experiment. Rectangles mark the OptiTrack cameras, and hexagons mark the HTC VIVE cameras. A loudspeaker (Behrtione C50A from Behringer[8]) stands at a slightly off-center position in the room and is marked with circles in the same figure.

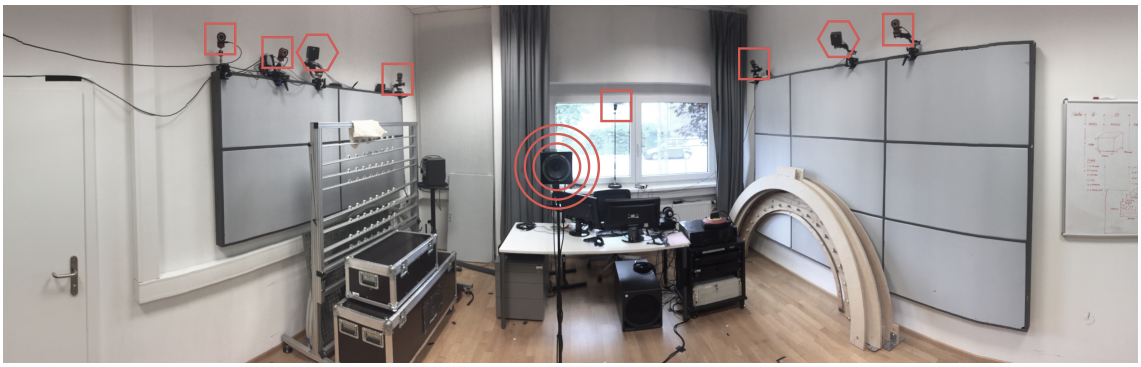


Figure 2.1 – The position of the real loudspeaker is the same as in the virtual studio and is marked with red circles. Red rectangles identify OptiTrack cameras, and red hexagons indicate the two HTC VIVE cameras.

Section 1.4 explained why it was necessary to implement two parallel systems for the tracking of the participants. Figure 2.2 shows both systems. The green connections mark the HTC VIVE head-mounted display and the VR controller. The HTC VIVE System uses infrared light bursts to calculate the position of the HMD and the controllers in relation to the tracking cameras. The red connection marks the OptiTrack system. The small white balls on the top of the head of the participant form one tracked object. This object was mounted on the headphones because they were used even when the HMD was taken off to see the real room. The LEDs that are used by the OptiTrack cameras emit 850nm IR light, which is reflected by the small balls. This allows the system to precisely track the position and orientation of the object in the room. The OptiTrack is a measurement system, which means it is more precise than the VIVE tracking. It would have been desirable only to use OptiTrack to provide Reaper with the orientation and position data of the participants.

8. `https://www.behringer.com`

Unfortunately, the two infrared systems interfere with each other, which results in strongly drifting and lagging VR visuals with the HMD. Therefore the VIVE was used whenever the participants did use the HMD, and OptiTrack was used whenever they did not.

Another difficulty with the tracking systems is the covered space in the room. Both the VIVE and the OptiTrack system have difficulties with tracking the movement of participants who use the whole room to move around, but the reasons for the difficulties differ. In case of the VIVE system, the tracking cameras have a field of view of 120 degrees and should not be mounted with a distance to each other of more than five meters [9]. After experimenting with the camera positions, it resulted that a setup with one station on the left and one on the right wall is best in this situation. With this tracking system, the room should not be bigger than the studio if the participants are allowed to use the whole room. The problem with OptiTrack is not the distance from the cameras to each other but their rather small field of view [10] with only 56°. In addition to that, at least three cameras have to identify the tracked object at any time. This results in the need for many cameras, especially if the participants want to walk all the way up to the walls of the room.

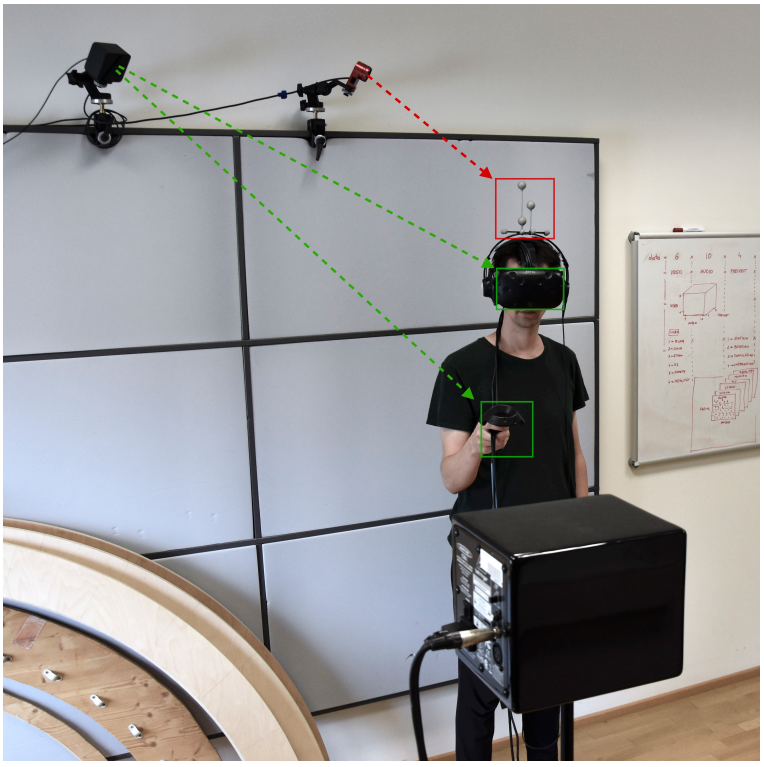


Figure 2.2 – The setup for the participants during the experiment. Red = Optitrack system; Green = HTC VIVE system. Both systems were used to gather position and orientation data. The OptiTrack system was used only when the participants did not wear the HMD. In front of the participant, one can see the real loudspeaker.

9. https://www.vive.com/us/support/vive-pro-hmd/category_howto/tips-for-setting-up-the-base-stations.html
10. https://optitrack.com/public/documents/Flex%2013%20Data%20Sheet.pdf

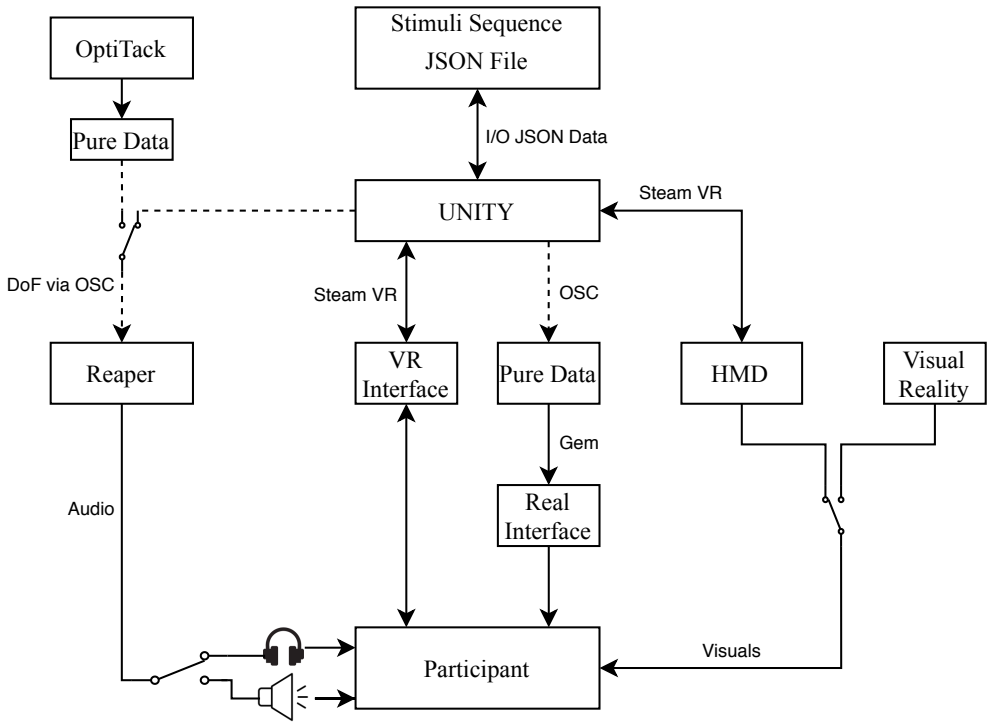



Figure 2.3 – The figure shows a scheme of the technical setup for the listening experiment.

The core of this listening experiment was the software "Unity." On the one hand, Unity was providing the visuals to the HMD via the software "Steam VR." On the other hand, it was managing the procedure of the listening experiment. To be able to do so, it read from a JSON file which stimuli should be presented next and accordingly switched the virtual environments. It also sent the orientation and position data of the participant via open sound control (OSC) to Reaper. Reaper then modeled the acoustics of the virtual room, and Unity wrote the rating for all the stimuli back to the same JSON file. Figure 2.3 shows how the technical system was set up and how the individual parts were connected. These individual blocks will be discussed now in more detail. The modeling of virtual environments and the acoustic modeling will be discussed separately in chapters 3 and 4. For now, it is important to keep in mind that there were several different visual stimuli, including the real room, and several different acoustic stimuli, presented either via headphones or via a real loudspeaker.

### The VIVE VR system with Unity

The visual presentation of the virtual environments was done with the VIVE VR system. This system included one head-mounted display, two VR controllers, and two tracking cameras (also called "base stations"). Through the HMD, the participants could see the virtual reality that was provided by Unity. The link between Unity and the VIVE system

was a software called "Steam VR." Usually, Steam VR is used to download and play games for VR. In this case, it displayed the content Unity provided. To show a Unity scene via the HMD, one has to load the Steam VR Asset from the Unity Asset Store [11] and import it to the Unity project. To see the modeled virtual environments with the HMD, one has to use the "CameraRig" provided in SteamVR/Prefabs.

For the listening experiment, it was important that the real and the virtual rooms were lined up. If that wasn't the case, the virtual and the real loudspeaker would not have been in the same position in relation to the participants. To be able to aline real and virtual environments, it is necessary to understand how the calibration of the VIVE system works and how it affects the position of the virtual environment. The calibration of the VIVE system is done in Steam VR and it is called "Room Setup." To aline the virtual and the real room, it is best to use the "Set up for Standing Only" - option during the calibration. With this option, SteamVR will ask the user to position the HMD in the center of the playing area. If the HMD is put on the floor in the center of the room that has been virtualized, the virtual room will have its origin right there later. To prevent the virtual and the real room from having a different rotation, one has to make sure the HMD is looking in the forward direction during the calibration. The real and the virtual room will be aligned if the Unity scene has its coordinate origin in the center of the virtual room.

### Open Sound Control Connections

To render the modeled acoustics to a pair of headphones, the digital audio workstation Reaper needs information about the position and the orientation of the participant in VR. This information is provided to Reaper by Unity via the communication protocol "open sound control" (OSC) [12]. OSC messages allow Unity to manage the Reaper session flexibly and to do things such as muting channels or changing parameters in Plug-ins. Thomas Fredericks free C#-scripts [13] did the OSC communication in Unity.

### Opti Track

When the participants did not use the HMD, a second tracking system had to be used to gather their position and orientation data. This system was the OptiTrack. Pure Data was used to convert the data stream from OptiTrack to OSC messages that were readable to Reaper.

11. `https://assetstore.unity.com/packages/tools/integration/steamvr-plugin-32647`
12. `http://http://opensoundcontrol.org/introduction-osc`
13. `https://thomasfredericks.github.io/UnityOSC/`

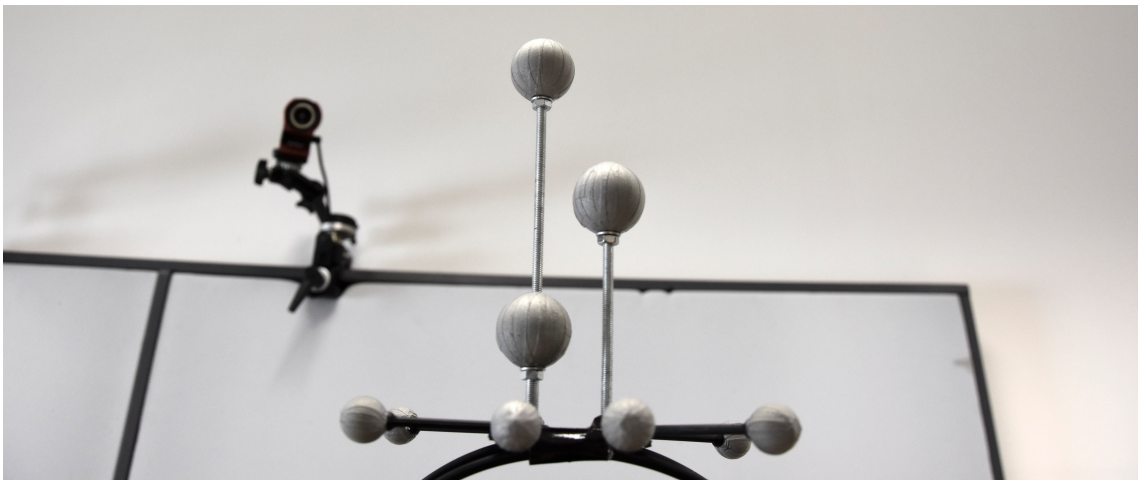


Figure 2.4 – The OptiTrack object was constructed of six small balls and three slightly larger balls. The structure was mounted on the top of the headphones of the participants.

### JSON

JSON [14] stands for JavaScript Object Notation and is a human-readable format to interchange any text-based data. On the one hand, the format was used to create the datasets containing the individual stimuli sequences for all of the participants. On the other hand, it was used to store the ratings of the participants. A hypothetical example shows how a file with only four stimuli would have been structured:

```
{
    "Reihenfolge": [
        3,
        1,
        4,
        2
    ],
    "Antworten": [
        2,
        7,
        6,
        2,
    ]
}
```

Listing 2.1 – This is the structure of an individual JSON text file for a participant. In the real experiment, these lists had not only four but 136 entries.

"Reihenfolge" is german for "sequence" and "Antworten" is german for "answers." In this case, the stimuli number three was presented first and the participant rated with a value of two. The second presented stimulus was stimulus number one, and the participant rated this one with a value of 7. The interpretations of these ratings will be explained in section 5.2.

14. https://www.json.org/json-en.html

## 2.2 Interface Design

To be able to rate the different stimuli, the participants needed to use some interface. The VR controller was used to control an interface in virtual reality and reality. How the interface was designed in terms of content will be explained in section 5.2, here the technical implementation is of interest. Whenever the participant was in virtual reality, the interface was also shown in virtual reality. A transparent canvas displayed a text in front of the participant. This canvas was attached to the CameraRig and therefore moved together with the vision of the participant. Rotating the VR controller let the participants choose their rating on a discrete scale from one to seven. For the stimuli that are done without the HMD also the interface display needed to change. Therefore in these cases, a monitor in the studio displayed the same text passages as the canvas in VR. Pure Data displayed the messages on the real monitor via its "Graphics Environment for Multimedia" (Gem [15]). Figure 2.6 shows how Pure Data received the VR controller data, and figure 2.7 shows the PD patch that displayed the text on a monitor via Gem.


Figure 2.5 – The user interface that was displayed on a transparent canvas in the virtual studio.

15. http://gem.iem.at/documentation/faq/what-is-gem

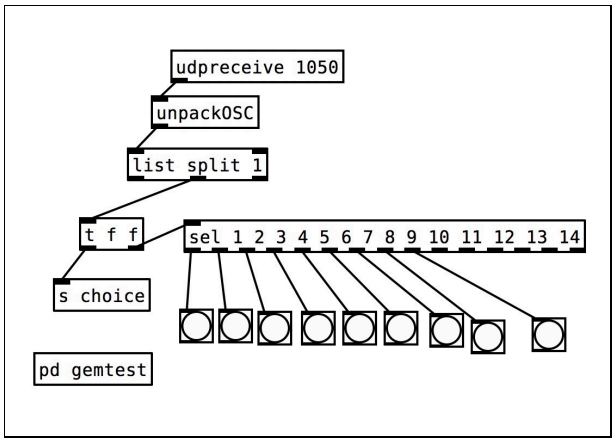



Figure 2.6 – Pure Data receiving OSC messages from Unity.

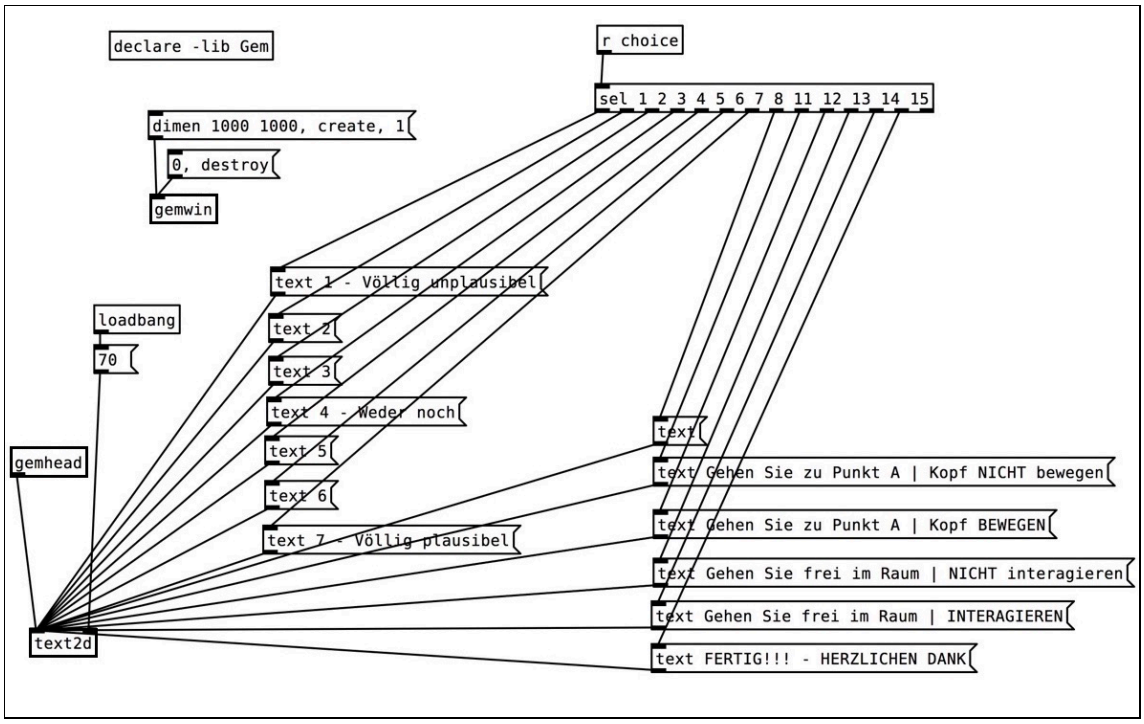



Figure 2.7 – The Pure Data patch that displayed the interface text on an external monitor.

# Chapter 3

# Visual Conditions

This chapter presents the five different visual virtualizations that were used in the listening experiment. All in all, the following six visual conditions were part of the experiment:

1. Studio
2. Dark Room
3. Shoebox
4. Lecture Room
5. Concert Hall
6. Reality

First, the software "Unity" is introduced.

## 3.1 Modeling with UNITY

Unity [1] originally was a game development engine but is nowadays used for other applications such as architecture, construction, and film. It offers the user to design 3D virtual environments in a very flexible way. In Unity, the development is done in C# - an object-based programming language. GameObjects are the objects which can be modified through the assignment of C# scripts. Basic functionality such as changing position and size of GameObjects can be done directly via the Unity interface. More complex functionalities such as sending OSC messages need to be scripted. The environments designed for this listening experiment are mainly done with the GameObjects "Cube" and "Cylinder," which were modified in size and shape. Figure 3.1 shows the Unity Editor. In the top-left part, one can see the scene from any desired perspective. Below that, one can see the rendering camera's view. This name of this camera is "Main Camera," and it is part of the "Hierarchy" in the top-center of the editor. Here all the GameObjects are listed and can be structured. Below that, one can see the "Project" folder, which contains all the assets, packages, and scripts necessary to render the scene.

1. `https://unity.com`

Last but not least, the "Inspector" on the right side of the window is responsible for setting all the parameters and structuring the C# scripts used for the individual GameObjects. In this figure, the Inspector displays the parameters of the cube. One can read from this section that the position of the cube is at the origin of the scene, it is not rotated in any direction, and its scale has not been changed. Figure 3.2 shows Unity's scene window in the process of modeling the concert hall.

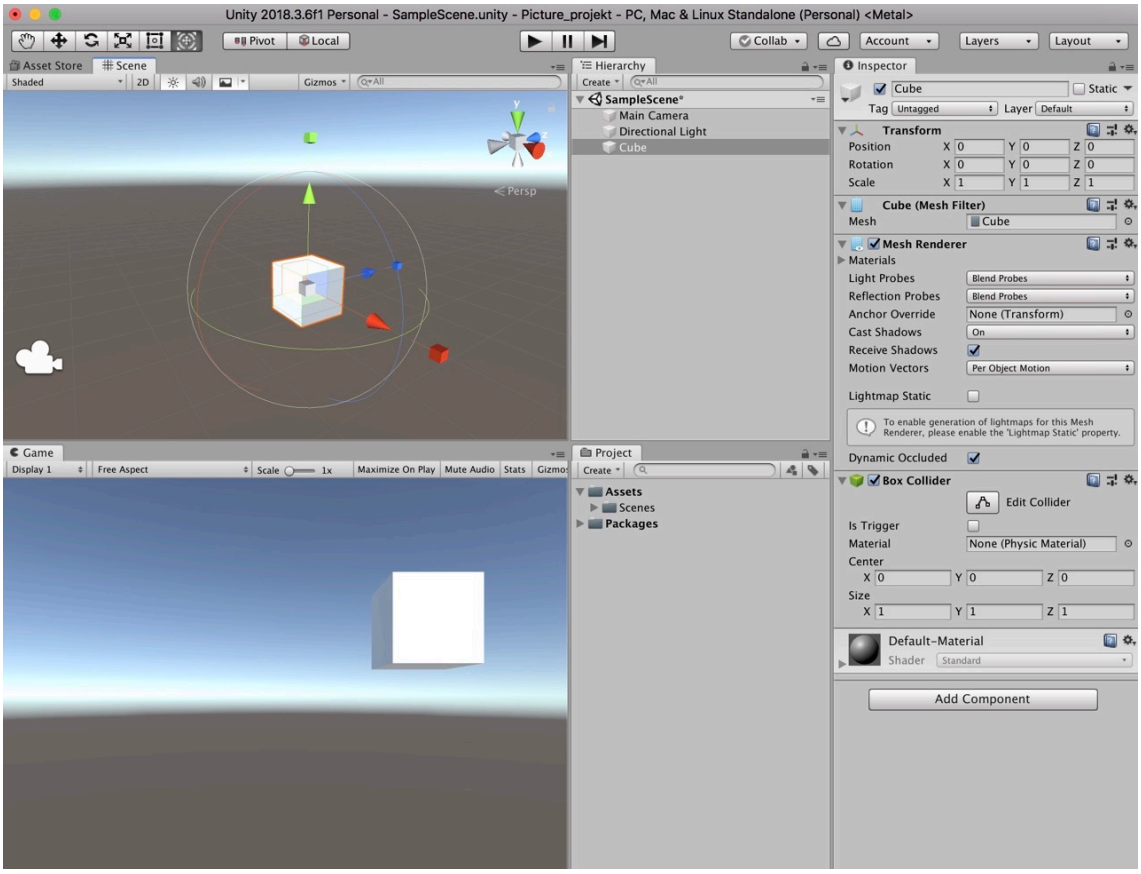



Figure 3.1 – A default scene in Unity with the Gameobject "Cube" (1m x 1m x 1m).

## 3.2 The visual conditions

### The studio

Figure 3.3 shows a picture of the real studio. Figure 3.4(a) shows the studio from a similar perspective. The goal of the virtualization is to match the real room in the most important aspects such as size, basic interior, light, and spatial relation to the user. To provide the participants with a visually more appealing display, a scene outside the window was recorded with a 360° camera and played back via the skybox in unity. A skybox represents the most distant boundary in a Unity scene - like the sky in reality. It is possible to project 360° pictures or videos on them. This way, the users had the same view outside of the

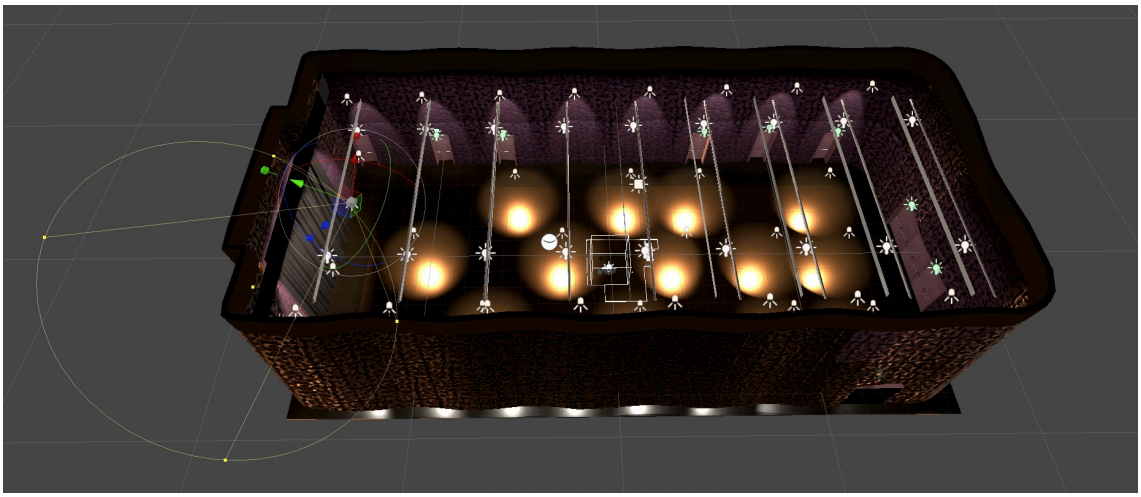


Figure 3.2 – The modeling of the concert hall - the virtual György-Ligeti-Haal. Here one of the spotlights is positioned to illuminate the stage curtain.

Figure 3.3 – The real studio from the perspective of the participants. The picture was taken from position 'A'.

(a) The studio from the perspective of the participants.

(b) The studio from an aerial perspective.

Figure 3.4 – The virtualization of the studio, designed with Unity.

window, with and without VR goggles. For information concerning the size of the room and the position of the loudspeaker, please see table 4.2. The volume of the room is about 84 $m^3$.

## The dark room

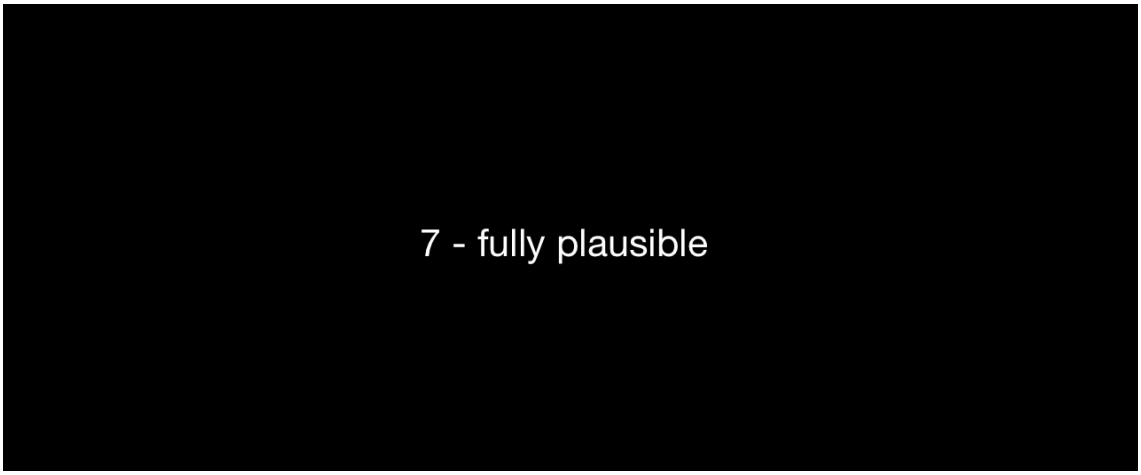



Figure 3.5 – In model of the the dark room only the interface was visible.

The dark room was designed as a simple adaption of the studio room: the virtual lights were turned off. To be able to see the user-interface, its letters were white instead of black in this room. Furthermore, it was necessary to be able to see the HTC VIVE controller. Therefore a virtual light source with a range of 0.3 meters was connected to the controller. This way, the controller was visible while the room remained dark. The goal of this visual condition was to see if the results match the ones of other research findings such as in [ZFZ18].

## The shoebox

The shoebox virtualization was equal to the studio virtualization except that in this one, there was no interior, there were no doors, and no window. From a technical point of view, this room is the perfect match for the acoustic model as also the Room Encoder (responsible for the discrete acoustic reflections) simulates a perfect shoebox model. Its purpose was to investigate the effect of interior and visual details on the perception of plausibility. One can see an area marked on the floor with a white line. This line marked the space the participants were able to use. If they went outside this area, they would collide with an object in the real room. This "Participant area" was marked in all the visible rooms except the studio. In the studio it was not necessary to mark is because the objects in the room were modeled and therefore visible.

## The lecture room

The lecture room is the virtualization of the lecture room at IEM. It is about 60 $m^2$ large and 3.1 m high, which leads to a volume of 186 $m^3$. Figure 3.7(b) also shows this room from an areal perspective. The wall on the right side consists mainly of windows. Therefore also in this room, a 360° video of the exterior was mapped to the skybox in Unity.

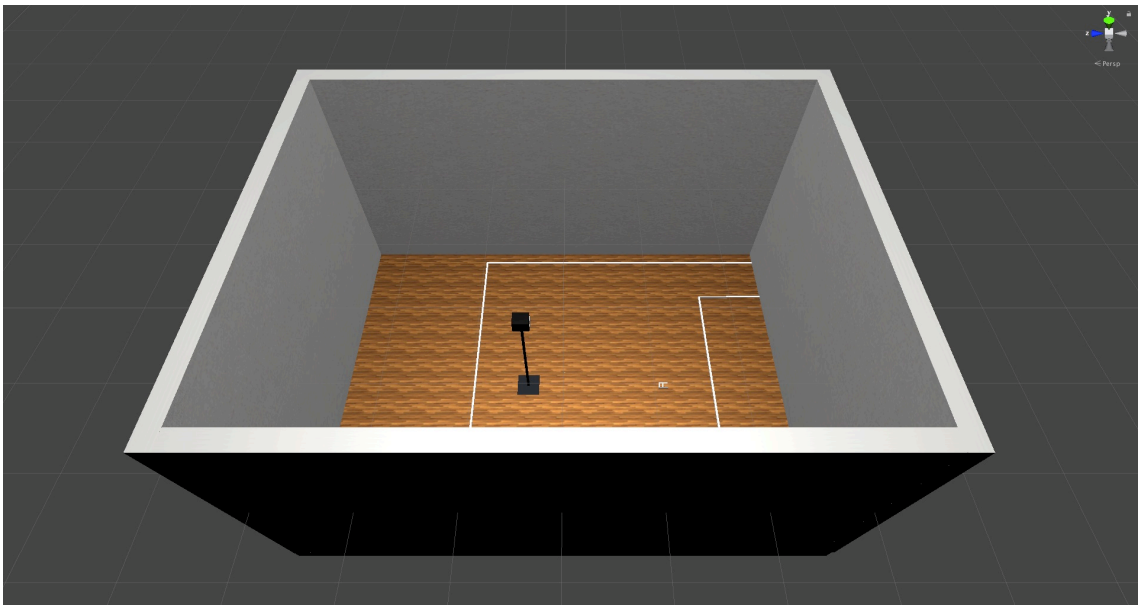


Figure 3.6 – The shoebox model. The room has the same measures as the studio but has no interior, no doors, and no window.

(a) The model of the lecture room from an areal perspective.

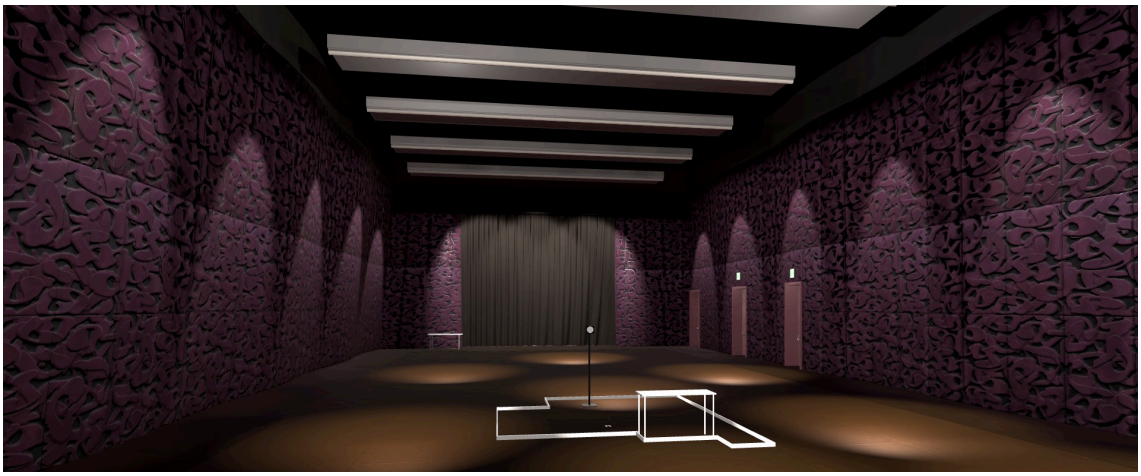
(b) The lecture room in reality.

Figure 3.7 – The lecture room at the IEM in its virtual and real form.

## The concert hall

The concert hall is the virtualization of the György-Ligeti-Haal at MUMUTH Graz [2]. It is about 500 $m^2$ large and 8 m high, which leads to a volume of 4000 $m^3$. Figures 3.8(a) and 3.8(b) show the virtualization of the concert hall and the perspective of the participants in it. The basic model of this hall was already provided; what's new is the light design in the room.

2. https://www.kug.ac.at/universitaet/campus-und-gebaeude/mumuth/



(a) The concert hall virtualisation.

(b) The concert hall from the perspective of the participants.

Figure 3.8 – The György-Ligeti-Haal.

# Chapter 4

# Acoustic Conditions

A total of nine different acoustic models and a real loudspeaker were examined in the listening test. The nine virtualizations were modeled using various techniques. At first, this chapter informs about several measurements that were necessary for the acoustic modeling. Then the different paths of the acoustic modeling are presented in a flowchart. Several tables show the parameter settings of the used plug-ins. Afterward, the creation of the 7th-order model is described in detail because it is the basis for all other virtualizations as well. Finally, the nine remaining acoustic conditions are explained. The ten different conditions were the following:

1. Real loudspeaker
2. 7th Order
3. 3rd Order
4. 1st Order
5. No Directivity
6. BRIR
7. BRIR & No Directivity
8. Short Reverb
9. Lecture Room
10. Concert Hall

## 4.1 Required acoustic measurements

This section provides information about the measurements, which were necessary for the acoustic modeling.

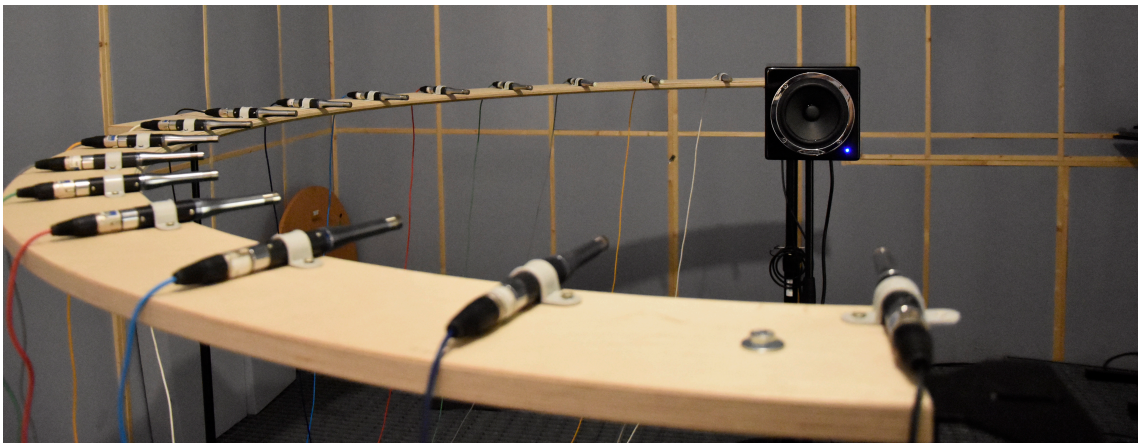


Figure 4.1 – Measurement of the loudspeaker directivity pattern with 17 microphones.

| Mic Nr. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Azimuth** | 0° | 11.25° | 22.5° | 33.75° | 45° | 56.25° | 67.5° | 78.75° | 90° | 101.25° | 112.5° | 123.75° | 145° | 146.25° | 157.5° | 168.75° | 180° |
| **Elevation** | 0° | 0° | 0° | 0° | 0° | 0° | 0° | 0° | 0° | 0° | 0° | 0° | 0° | 0° | 0° | 0° | 0° |

Table 4.1 – The 17 microphone positions for the measurement of the directivity pattern of the loudspeaker. The angle between two consecutive microphones is always 11.25°.

### Loudspeaker Directivity Pattern:

One part of the quality of plausible acoustic modeling certainly is a proper directivity pattern of the virtual loudspeaker. The hypothesis is that a directivity pattern will improve the plausibility of the model. To be able to design the correct directivity pattern, the true pattern first had to be measured. The Behritone C50A loudspeaker is symmetric around his main axis and uses only one driver. It is reasonable that also its directivity pattern is symmetric around this axis. The measurement was done in an anechoic chamber with 17 microphones at a distance of one meter to the loudspeaker. The microphones were set up in a semicircle around half of the loudspeaker at the positions from table 4.1. Due to the axially symmetric design, all of the relevant pattern-information is contained in a measurement at these positions. Figure 4.1 shows the measurement setup. In the measurement, a sweep from 50 Hz to 22050 Hz was played back via the loudspeaker resulting in a 17-channel WAV file. The 17 impulses were deconvoluted with equation 4.1. According to Pythagoras, the first reflection arrives from the floor after about 215 samples. The direct sound was cut out of the impulse signals with a Tukey window with r = 0.1 [Blo04]. How the measurements were implemented in a virtual loudspeaker directivity pattern will be described in section 4.2.

$$h(t) = IDFT\left[\frac{DFT(y(t))}{DFT(x(t))}\right], \tag{4.1}$$

Where:

| | |
|---|---|
| $x(t)$: | is the original sweep signal. |
| $y(t)$: | is the recorded sweep signal at the output of the system. |
| $h(t)$: | is the deconvoluted impulse response of the system. |
| $(I)DFT$: | (Inverse) Discrete Fourier Transformation. |

#### Binaural Room Impulse Response:

One of the acoustic conditions in the experiment used a BRIR measurement to model the reverberation. To be able to do so, a BRIR was measured with the Neumann KU100 artificial head. The artificial heads position during the measurement was position 'A', hence the same position at which also the participants were standing when they were not allowed to move their bodies. The height of the ears of the KU100 was 1.6 meters, and the distance to the loudspeaker was 1.8 meters. The KU100 and the Behringer loudspeaker faced each other. For the measurement of the BRIR, a sweep from 50 Hz to 22050 Hz was played back. The sweep lasted for 15 seconds. Again equation 4.1 was used to obtain an impulse signal for the left and one for the right ear. The BRIR measurement was used for two purposes: On one hand, it was used to design the static BRIR reverb. On the other hand, it was used in the process of designing the parametric reverb in the virtual studio.s

#### Impulse response of the open headphones:

Since one part of the listening experiment was to listen to the real loudspeaker, it was crucial to be able to hear the acoustics of the real room through the headphones. Therefore special open headphones were used [MKRB$^+$20]. The modifications on these headphones influence its sound, so it was necessary to compensate this with an extra filter. How this filter was designed will be explained later. In the measurement, the KU100 wore the pair of open headphones, which played back a sine sweep from 50 Hz to 22050 Hz that lasted for 15 seconds. Once again, equation 4.1 was used to obtain the impulse response of the open headphones for the KU100.

## 4.2 Acoustic modelling

Prior to the detailed explanation of the 7th-order model, the overall system of the virtual acoustics setup is introduced here. Figure 4.2 shows a flowchart for the acoustic modeling. All the steps from this flowchart were done with the digital audio workstation "Reaper" [1], the "IEM Plug-in Suite" [2], and the multichannel audio plug-in suite by

1. www.reaper.fm
2. https://plugins.iem.at

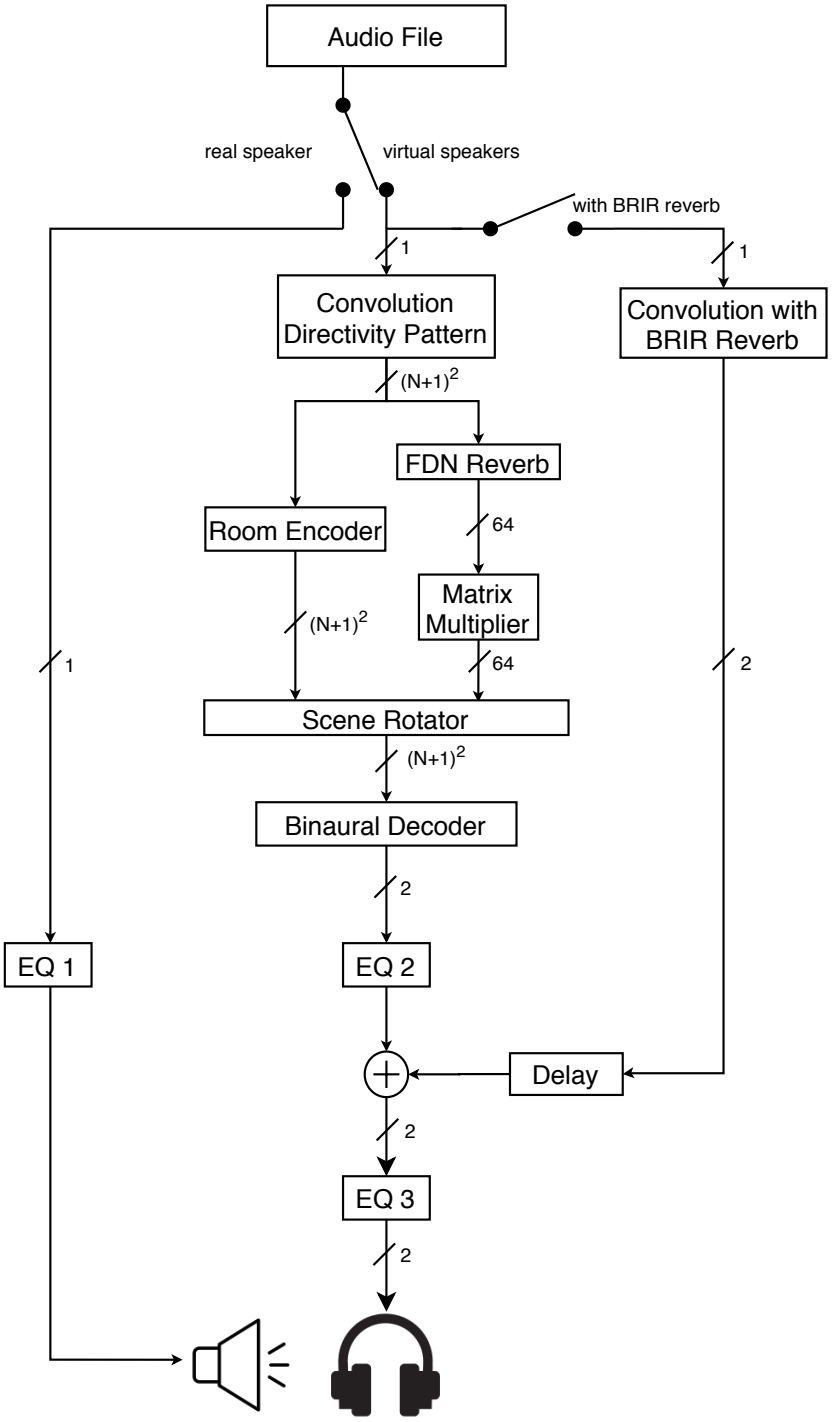



Figure 4.2 – The audio flowchart.

Matthias Kronlachner[3]. These Plug-ins offer a variety of VST Plug-ins to encode, manipulate, and decode Ambisonic signals. The essential blocks in the modeling were the "Room Encoder," the "FDN Reverb," the "Scene Rotator," and the "Binaural Decoder."

3. http://www.matthiaskronlachner.com

| Room | # Reflections | Room Dimensions in m | | | Wall Attenuations in dB | | | | | | Overall Attenuation |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | X | Y | Z | Front | Back | Left | Right | Ceiling | Floor | |
| Studio | 236 | 6 | 4.4 | 3.2 | -3 | -9 | -3 | -3 | 0 | 0 | 0 |
| Lecture Room | 236 | 7.3 | 8.4 | 3.1 | 0 | -3 | -2 | 0 | 0 | 0 | 0 |
| Concert Hall | 236 | 30 | 15 | 9 | -15 | -8 | 0 | 0 | -2 | 0 | -5 |
| Short Reverb | 236 | 6 | 4.4 | 3.2 | -3 | -9 | -3 | -3 | 0 | 0 | -6 |
| BRIR Reverb | 0 | 6 | 4.4 | 3.2 | - | - | - | - | - | - | 0 |
| No Directivity | 236 | 6 | 4.4 | 3.2 | -3 | -9 | -3 | -3 | 0 | 0 | -4 |

Table 4.2 – Parameter settings of the Room Encoder Plug-In for the different room situations. The low shelf was set to 283 Hz with a gain of 4.5 dB for all the rooms. The high shelf's gain were set to 0.

| Room | Room Size | Reverb Time | Fade In Time | Highshelf Gain |
|---|---|---|---|---|
| Studio | 2 | 0.3 | 0.29 | -69 |
| Lecture Room | 2 | 0.5 | 0.49 | -69 |
| Concert Hall | 6 | 1.1 | 1.09 | -10 |
| Short Reverb | 2 | 0.15 | 0.14 | -69 |

Table 4.3 – FDN Parameters for the different room situations. The low shelf was always set to 219 Hz with Q = 0.73 and Gain = 4.1 dB. The high shelf was set to 1100 Hz with a Q of 0.9 and a gain according to the table.

The Room Encoder was responsible for encoding the direct path and the first 236 reflections of the signal. These 236 reflections were modeled with image sources. The late and diffuse reverberation was modeled with the FDN reverb - a 64 channel feedback delay network. The Scene Rotator was in charge of rotating the Ambisonic scene according to the head movements of the participants to provide them with a stable sound field. In the end, the Binaural Decoder decoded these signals to the pair of open headphones. Figure 4.2 shows two more possibilities next to the central path. On the left side, it shows the path directly to the real loudspeaker and on the right side, it shows the path through a convolution with the reverberant part of the static binaural room impulse response. Whenever the real loudspeaker played back the signal, the modeling was muted. Whenever the BRIR Reverb was convolved with the signal, the FDN reverb was muted, and the Room Encoder only encoded the direct path of the sound. There are three tables providing information about the parameter settings of the Plug-ins. Appendix C provides the GUIs of the most relevant Plug-ins. Table 4.2 shows the settings of the Room Encoder for the acoustic conditions "Studio," "Lecture Room," "Concert Hall," "Short Reverb," "BRIR Reverb," and the "No Directivity" model. Table 4.3 shows the parameter settings of the FDN reverb Plug-in for the acoustic conditions "Studio," "Lecture Room," "Concert Hall," and "Short Reverb." In accordance with figure 4.2, table 4.4 gives an overview of the settings of the different blocks in the flowchart. It provides information about the activation of the real loudspeaker, the overall Ambisonics modeling order, whether the directivity pattern of the virtual loudspeaker is active or not, about the Room Encoder model, the FDN model, and whether the BRIR convolution is active or not.

| Conditions | Real Speaker | Ambi Order | Directivity | Room Encoder | FDN | BRIR Rev. |
|---|---|---|---|---|---|---|
| Reality | On | - | - | - | - | - |
| 7th Order | Off | 7th | Yes | Studio | Studio | Off |
| 3rd Order | Off | 3rd | Yes | Studio | Studio | Off |
| 1st Order | Off | 1st | Yes | Studio | Studio | Off |
| BRIR | Off | 7th | Yes | BRIR | Mute | On |
| No Directivity | Off | 7th | Off | Studio | No Directivity | Off |
| BRIR & No Dir. | Off | 7th | Off | BRIR | Mute | On |
| Short Reverb | Off | 7th | Yes | Short Reverb | Short Reverb | Off |
| Lecture Room | Off | 7th | Yes | Lecture Room | Lecture Room | Off |
| Concert Hall | Off | 7th | Yes | Concert Hall | Concert Hall | Off |

Table 4.4 – This table shows the settings of the different blocks from figure 4.2. Within the columns "Room Encoder" and "FDN" the entries refer to the tables 4.2 and 4.3.

## 7th-order Ambisonics model

The 7th-order model is a fully dynamic model of the acoustics of the studio. It uses only the central path through the flowchart from figure 4.2. For the 7th-order model, the Room Encoder and the FDN reverb used the settings from the tables 4.2 and 4.3. The creation of the 7th-order model will be discussed along the central path of figure 4.2. First, the virtual directivity pattern of the loudspeaker needed to be implemented.

### Modeling the directivity pattern

```
%% ENCODING to Ambisonics Directivity Pattern

load impulse.mat;          % [252x17] 17 mic-impulses
az = [0:180/16:180]';      % [17x1] azimuth angles

% Find order gains via unassoziated Legendre polynomials
N  = 3;
m  = 0:N;
P  = legendre_u (N, cosd(az)) * diag (sqrt(2 * m + 1)/sqrt(2));
wn = pinv(P) * impulse';
WN =  [ wn(1,:); 0th order gain
        wn(2,:); wn(2,:); wn(2,:); % 1st order gains
        wn(3,:); wn(3,:); wn(3,:); wn(3,:);wn(3,:); % 2nd order gains
        wn(4,:); wn(4,:); wn(4,:); wn(4,:); wn(4,:); wn(4,:); wn(4,:)];

% 3rd-order beam to frontal direction
beam = getSH(N,[0 90*pi/180],'real')';
BEAM = repmat(beam,1,252);

% Encoding to an Ambisonics directivity pattern
IMP_encoded = WN.*BEAM;
```

Listing 4.1 – Matlab code for the encoding of the loudspeaker directivity pattern

The design process of the directivity pattern of the virtual loudspeaker was strongly dependent on the assumption that the pattern is axis-symmetric. This way, it was possible to use the measurements of only the 17 microphone positions on the semi-circle in figure 4.1. Listing 4.1 shows the Matlab-Code used to design the pattern.

The goal is to find order gains that shape a beam in the forward direction according to the measured directivity pattern. The shapes of Ambisonics beams are only influenced by the relations between the orders of the used Spherical Harmonics, not by the relations between the degrees. Usually the order-weights $w_N$ in equation 4.2 are used to suppress side-lobes of the Ambisonics signal [ZF19]. In general this equation decodes an Ambisonics signal $\boldsymbol{\chi}_N$ with the order gains $\boldsymbol{w}_N = [w_0, w_1, w_1, w_1, w_2, w_2, w_2, ..., w_N]^T$ to the loudspeaker signals $\boldsymbol{x}$ via the decoder matrix $\boldsymbol{D}$:

$$\boldsymbol{x} = \boldsymbol{D}\ \mathrm{diag}\{\boldsymbol{w}_N\}\ \boldsymbol{\chi}_N \tag{4.2}$$

In this case, the weights $w_N$ did not suppress side-lobes but shaped a beam to match the beam of the real loudspeaker. The decoding was done at a later stage with a binaural decoder. The Spherical Harmonics in the central column of figure 1.2 are independent of the azimuth angle. Hence they are axis-symmetric. These Spherical Harmonics have degree 0 and, therefore, only the unassociated Legendre Polynoms (see equations 1.2 and 1.3) influence their values on the sphere. It is necessary to sample axis-symmetric Spherical Harmonics at the angles of the 17 microphones to obtain the order gains required for modeling a directivity pattern in the z-direction. These order gains were obtained by evaluating the unassociated Legendre Polynomials (= Spherical Harmonics for degree 0) for 0th, 1st, 2nd, and 3rd order at the 17 microphone positions. For this purpose, the angles of the measurements had to be interpreted to be zenith angles. As the shape of an Ambisonics beam is just dependent on the gain relations within the orders, we can now use the order gains we found for the z-direction to shape a beam towards the frontal direction.

A directivity pattern with higher Ambisonics order did not result in a better representation of the real pattern. Frank and Brandner also show in [FB19] that for a speech signal, participants were not able to distinguish between the reference, a 7th-order model, and a 3rd-order model. The pattern-differences between the real and the virtual loudspeaker were investigated by comparing the figures 4.4 and 4.3. To be able to see the virtual directivity pattern, it was decoded at the 17 microphone angles. These 17 signals were transformed into the frequency domain and visualized in the same way as the real recordings. During the listening experiment, the "MCFX-Convolver" - one of Matthias Kronlachner's Plug-ins - was used to convolve the mono WAV file with the 16 impulse responses to obtain a 3rd-order Ambisonics signal. After the MCFX-Convolver, the 16 channels were sent to the Room Encoder and the FDN reverb.

### Room Encoder

The input of the Room Encoder always used the 3rd-order directivity pattern of the loudspeaker, regardless of the model order. The output used the 7th order in this case. The output order changed for the 3rd and the 1st order model. The Room Encoder was respon-

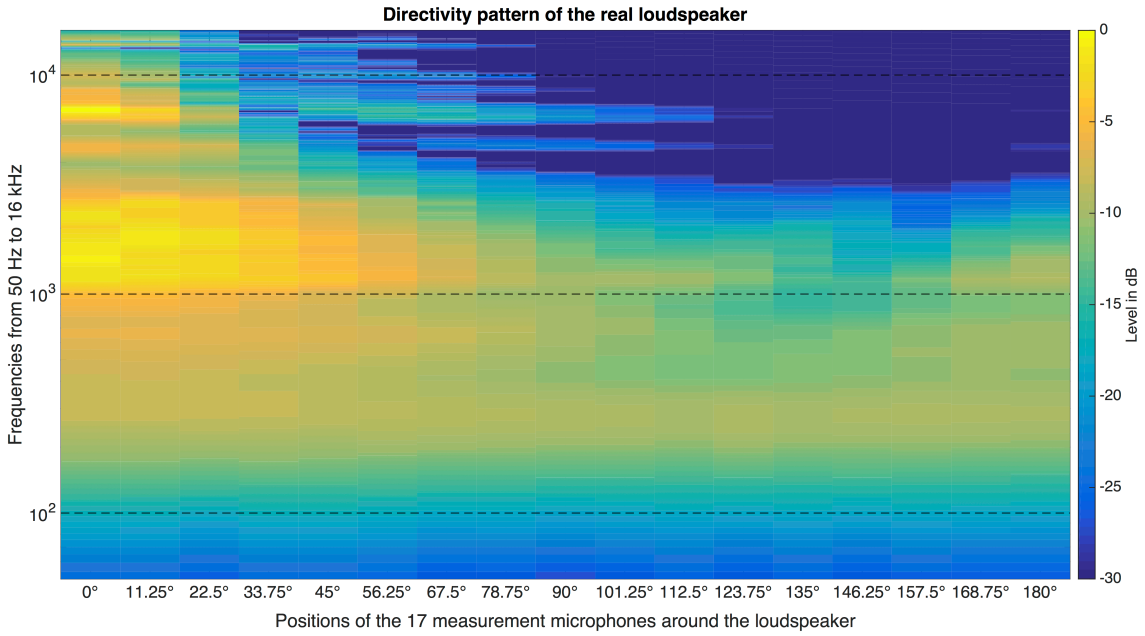



Figure 4.3 – The directivity pattern of the real loudspeaker. The figure shows a spatially dependent spectrogram over the 17 microphone positions from 0° (front) to 180° (back). The colormap shows levels between -30 dB and 0 dB. Frequencies range from 50 Hz up to 16kHz.

sible for the encoding of the early reflections and the direct sound. It generated 236 early reflections of the virtual loudspeaker by calculating them with an image-source-model. The Plug-in allows changing the damping factor of all the walls, the ceiling, and the floor. These parameters were set to the values in the row "Studio" from table 4.2. The Room Encoder also delays the direct path of the sound according to the distance of the participant to the virtual loudspeaker. This way, the Plug-in also considers the propagation speed of real sound. The position of the loudspeaker was set to [x = 0.4; y = 0.2; z = 0], just like in the real room. Note that the origin in the Room Encoder is in the center of the room. Therefore this z-value resulted in a loudspeaker with a height of 1.6m.

### FDN Reverb

While the Room Encoder created the early reflections, the FDN reverb was responsible for the late and diffuse reverberation. The sound of an FDN reverb changes with the number of feedback channels it uses. Therefore this FDN used a fixed amount of 64 channels. These 64 channels then had to be encoded to a 7th-order Ambisonics reverb via multiplication with a $64 \times 64$ encoding matrix, to provide an order-independent impression of the decay length. The 7th-order model would have sounded far more reverberant than the 1st-order model if the 64 channel FDN output was not encoded to an Ambisonics signal. The encoding matrix contained the 64 values of the Spherical Harmonics evaluated at 64 almost equally-spaced positions. It is a coincidence that the Hadamard Matrix in

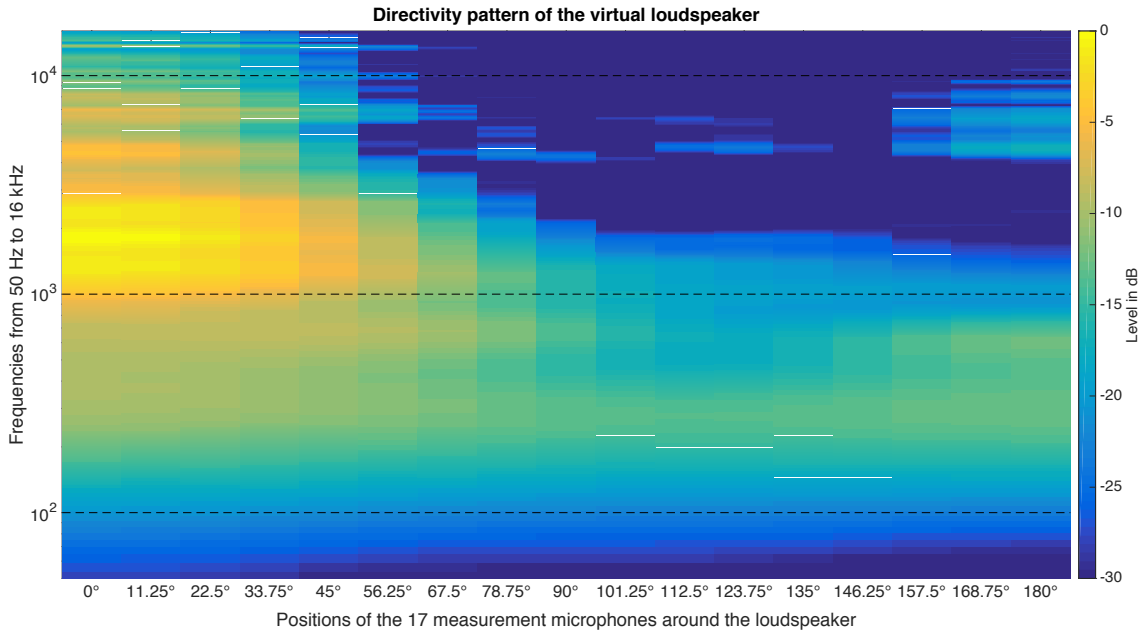



Figure 4.4 – The directivity pattern of the virtual loudspeaker. The figure shows a spatially dependent spectrogram over the 17 microphone positions from 0° (front) to 180° (back). The colormap shows levels between -30 dB and 0 dB. Frequencies range from 50 Hz up to 16kHz.

the feedback path of the FDN uses 64 channels [$2^6 = 64$] and 7th-order Ambisonics also requires 64 channels [$(N+1)^2 = (7+1)^2 = 64$]. The settings of the FDN reverb for the studio can be found in table 4.3.

The reverb in the virtual room was modeled by the Room Encoder and the FDN reverb. To set both Plug-ins with reasonable parameter settings, a comparison was conducted between the measured BRIR at position 'A', and the virtual BRIR at the same position. To be able to see the influence of the Room Encoder and the FDN reverb on the overall reverb, the outputs of those two Plug-ins were rendered separately. Then the BRIR measurement, the Room Encoder output, and the FDN output were plotted against each other. This way, the reverberation in both Plug-ins were set to match the BRIR measurements as good as possible. Figure 4.5 shows how the comparison looked during the design process.

It has turned out that measurements and mathematical considerations can approximate the modeling well, but in the last step, the human ear is a good instrument for fine-tuning. Throughout the entire process of modeling, the ears were constantly used to check how the similarity to the real room was developing. In the end, the final models sounded a bit too reverberant in comparison to the real loudspeaker. Therefore the decay was adapted to match the impression of the real loudspeaker in the real room.

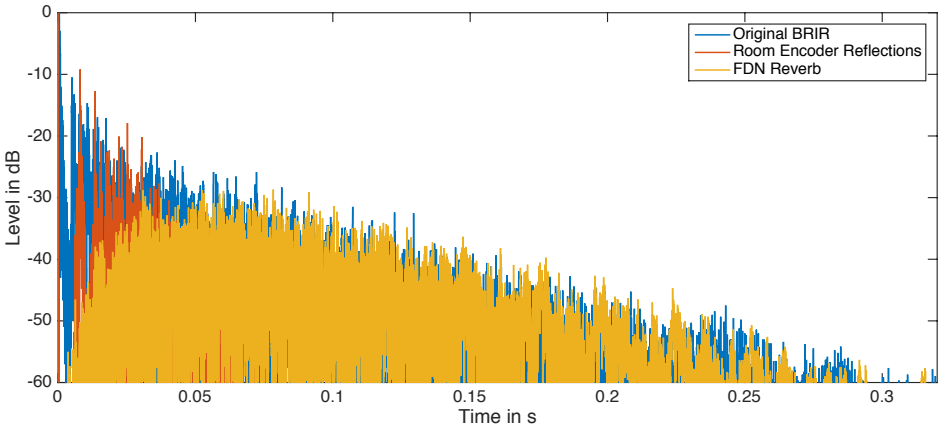



Figure 4.5 – The decay length of the Room Encoder and the FDN reverb were set while comparing the individual outputs to the measured BRIR.

### Scene Rotator

After the parallel processing of the Room Encoder and the FDN reverb, both 7th-order signals entered the Scene Rotator Plug-in for the next step. The Scene Rotator received information about the participant's position and orientation (6 DoF) from Unity via OSC. The scene was rotated according to the movements of the participants to provide them with a stable sound field. As there was a small delay audible for very fast head-rotations, the participants were asked to move rather slowly. The rotation of Ambisonics signals is described in [PH07, ZF19].

### Binaural Decoder

After the rotation of the scene, the signal entered the decoder. In this case, the decoding was done binaurally with the help of the "Binaural Decoder". Decoding Ambisonics signals to binaural signals can be done with different techniques. The Plug-in uses the state-of-the-art technique described in [ZSH18]. The Plug-in also offers headphone equalization for several headphone models. Anyway, this study was conducted with special open headphones, which are not selectable in the Plug-in. Therefore the equalization of these special headphones was done separately, immediately after the binaural decoding with a convolution with an extra filter.

### Open Headphones Equalization

The open headphones used in this experiment poorly reproduce low frequencies, which needed to be compensated. Therefore a filter was designed that was convoluted with the binaural signal. To be able to do so, the impulse responses of the open headphones, which were measured with the KU100 earlier, were transformed into the frequency domain. There, the magnitude spectrum was smoothed, inverted, and lowpass filtered (4th-order

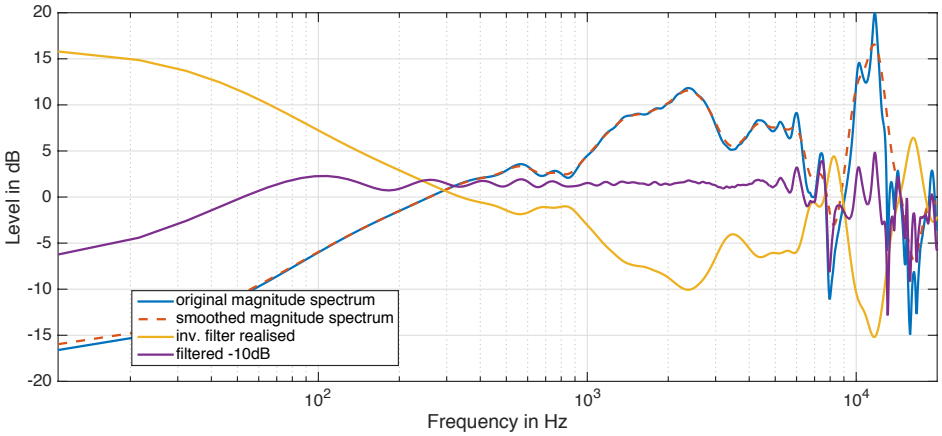



Figure 4.6 – The process of designing a compensation filter for the lack of low frequencies of the open headphones.

Butterworth filter at 18 kHz). This inverted magnitude spectrum was then transformed into an impulse response with the Matlab command `fir2`. The function `rceps` was used to obtain a minimum-phase impulse response. In the experiment, the headphone signal was convoluted with this minimum-phase impulse to compensate for the lack of low frequencies of the open headphones. Unfortunately, the convolution was done only with the path for the Ambisonics models in the Reaper session. Figure 4.2 shows with EQ2 that the compensation filter was missing in the path responsible for the BRIR reverb models. On the other hand, EQ3 was increasing the low frequencies for both paths in the BRIR models. EQ3 was a fine-tuning filter at the end of the signal chain.

## The remaining acoustic models

The eight remaining acoustic virtualizations are all related to the 7th-order model, and therefore they will be explained in relation to it. For details relating the parameter settings the tables 4.2, 4.3, and 4.4 can be used.

### 3rd and 1st-order models

In terms of content, the 3rd and 1st-order models work just like the 7th-order model. The parameter settings in all the Plug-ins are the same. The difference is the Ambisonics order they used to render the acoustic scenes. Figure 4.2 shows the different Plug-ins used in Reaper. The Plug-ins directly effected by changing orders were the Room Encoder, the Scene Rotator, and the Binaural Decoder. The FDN reverb is not per se an Ambisonics Plug-in, its output was always encoded with 7th order. The binaural decoding, in the end, was done with 3rd or 1st order, which affected the number of encoded Ambisonics channels but not the number of FDN output channels.

### Binaural Room Impulse Response (BRIR)

This model used only the direct path of the 7th-order model. This was done by muting the FDN reverb and by setting the number of discrete reflections in the Room Encoder to zero, compare 4.2. The directivity pattern for the virtual loudspeaker was, however, active in this model. Still, some reverb needed to be modeled, and this was done via a convolution of the speech signal with the reverberant part of the earlier measured BRIR in the real room. Section 4.1 explains how this measurement was done. As this convolution should only model the reverb, the direct path of the BRIR was removed and replaced with silence (210 samples). This file was convolved with the speech signal and delayed yet another time by 231 samples. The delay by 231 samples compensated for the delay the Room Encoder contributed during the encoding of the direct signal. The 210 samples of silence made sure the first reflection came at the right time after the direct sound.

The convolution of the speech signal with the BRIR reverb did lead to a static binaural reverb that was independent of the position and orientation of the participants. Therefore this condition is a combination of the 7th-order directivity-dependent direct sound and the static BRIR reverb.

The figures 4.7 and 4.8 show the setup during the BRIR measurement and when the participants stood on position 'A'. Using $D = 3,67\text{m} - 1,8\text{m}$ in equation 4.3 results in a delay of 240 samples instead of 210. During the measurement of the BRIR, the distances might not have been exactly like the ones in figure 4.7, the length of the silence in the BRIR reverb was measured directly from the BRIR file.

$$d = \frac{D}{343 * fs} \tag{4.3}$$

Where:

- $d$: is the delay in samples.
- $D$: is the distance in meters the sound needed to travel at 20 °C.
- $fs$: is the sampling frequency.

### 7th Order Without Directivity

This condition worked just like the 7th-order model with the difference that the loudspeaker had no directivity pattern and therefore emitted its sound the same way in all directions. Additionally, all of the discrete reflections from the Room Encoder were attenuated by 4 dB to prevent the condition from getting far more reverberant than the classic 7th-order model. Formally, "No Directivity" is an omni directivity pattern, and its implementation was done just like the one of the real pattern. Instead of 17 different microphone signals, here, all 17 channels contained copies of the signal from microphone number 1 (front microphone, compare table 4.1). This resulted in a 16 channel Ambisonics signal with only the first channel being different from zero.

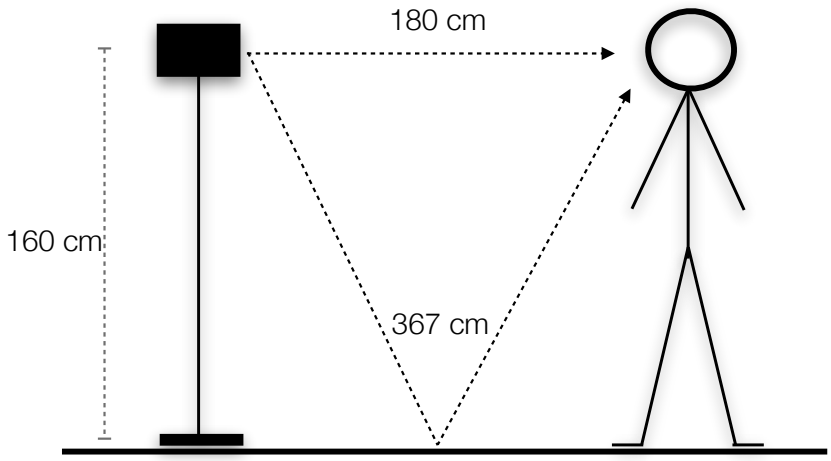



Figure 4.7 – When the participants stood on position 'A', they stood 1.8 meters in front of the loudspeaker. The path for the first reflection from the floor is 3.67 meters long. Both these paths needed to be compensated by a delay in the BRIR reverb convolution.

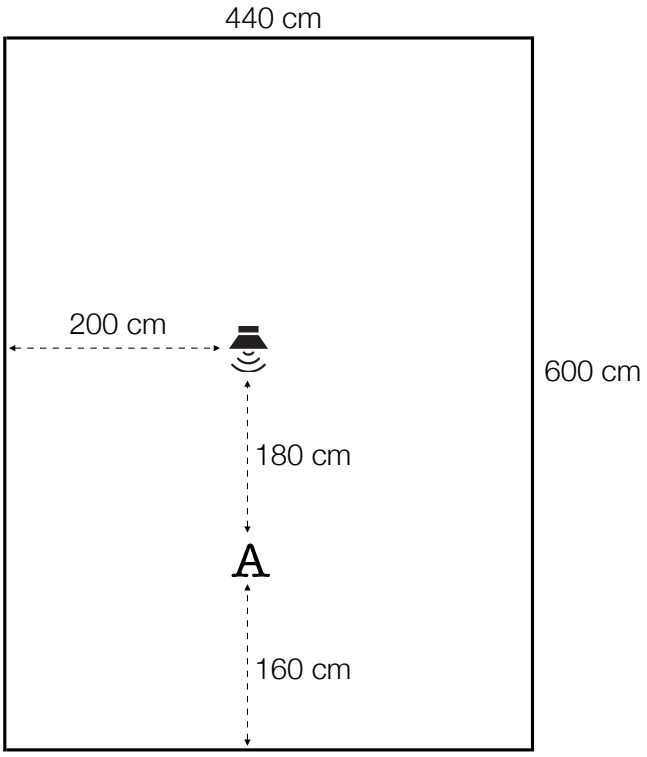



Figure 4.8 – The floor plan in the real and the virtual studio. The distance between the participant's default position 'A' and the loudspeaker is 180 cm. Their position is slightly off-center to make sure the discrete reflections in the acoustic model are not the same for the left and the right ear at position 'A'.

### 7th Order With Shorter Reverb

This condition also worked just like the 7th order, but the modeled reverb is much shorter here. This was realized with an attenuation of all reflections in the Room Encoder of 6 dB (compare table 4.2), and a reverb time of 0.15 seconds in the FDN reverb (compare table 4.3).

### BRIR & Without Directivity

This condition used the static BRIR reverb to model all the reflections. The direct sound was modeled with 7th order without loudspeaker directivity.

### Acoustic Lecture Room

The lecture room was modeled with the same technique as the 7th-order model of the studio. It used a reverberation time of 0.5 seconds and different attenuations of the walls. For details use tables 4.2, 4.3, and 4.4. The parameters were also estimated by comparing the virtual BRIR to a measured one from the lecture room.

### Acoustic Concert Hall

The concert hall acoustics were also modeled with the same technique as the 7th-order model of the studio. The parameter settings, in this case, were estimated by comparison to impulse responses from the concert hall. For details tables 4.2, 4.3, and 4.4 can be used.

### Real Loudspeaker

The open headphones and the HMD altered the sound of the real loudspeaker when it arrived at the ears of the participants. These influences were compensated by EQ1 from figure 4.2. Filter number two from figure 4.9 compensated the influence of the open headphones on the sound of the loudspeaker, and filter number one compensated the influence of the HMD. These filters were adjusted with the ears while taking on and off the headphones and the HMD. Automatic filter generation via a measurement was tested but didn't lead to a satisfying result. The amplification around 1320 Hz that was compensated by filter number two can also be found in a measurement of the open headphones in [MKRB$^+$20].

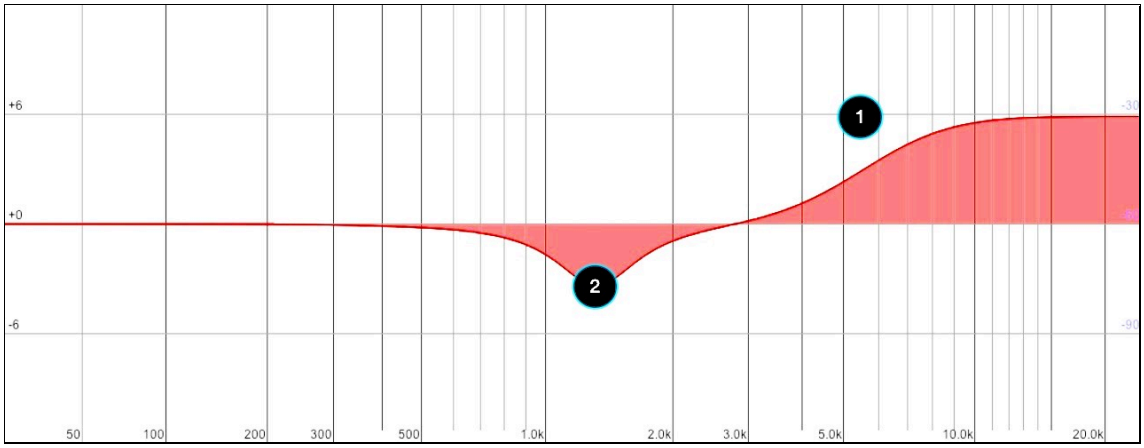



Figure 4.9 – The loudspeaker-participant transfer-path equalization filter. Filter 1: High Shelf, f1 = 5438.2 Hz, Gain1 = 5.9 dB, Bandwidth = 1. Filter 2: Band Filter, f2 = 1319.6 Hz, Gain 2 = -3.6 dB, Bandwidth = 0.8.

### Fine tuning the models

One of the last steps in the acoustic modeling was to compare the different conditions using the ears. This comparison resulted in two findings:

First, some of the conditions needed more level at low frequencies, and some needed less to obtain a uniform sound color for all the loudspeakers. Therefore, another filter was included just before the output to the loudspeakers. This filter was either set to the parameters from figure 4.10 or the ones from figure 4.11 to increase or attenuate the frequencies below 500 Hz. Table 4.5 shows which filter was used for which condition. Figure 4.2 shows the filter as "EQ3" at the end of the modeling path.

Secondly, the levels of the loudspeakers at the position 'A' were not all the same. The final step in acoustic modeling, consequently, was to make sure all the conditions were presented with the same loudness. To ensure this, the artificial head KU100 wore the open headphones while standing a position 'A'. The headphones played back the different models with white noise instead of speech. The signal recorded by the KU100 was then studied with a loudness meter[4]. According to the result of the loudness meter, one final gain equalized the loudness of the different conditions. Not only the headphone signals but also the signal of the real loudspeaker was taken into account here.

4. https://youlean.co/youlean-loudness-meter/

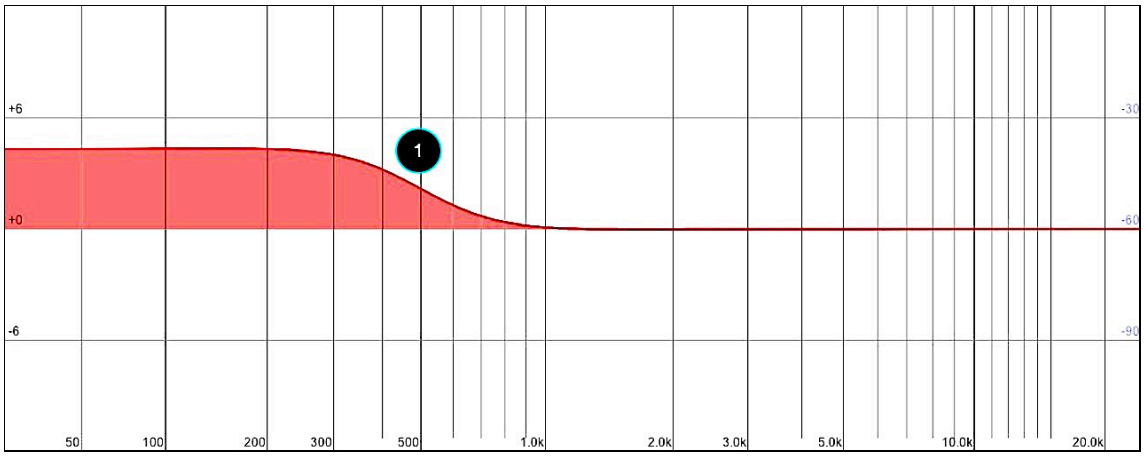



Figure 4.10 – EQ3: The fine tuning bass equalization filter for increasing low frequencies on the headphones. Filter 1: Low Shelf, f1 = 500 Hz, Gain1 = 4.3 dB, Bandwidth = 0.8.

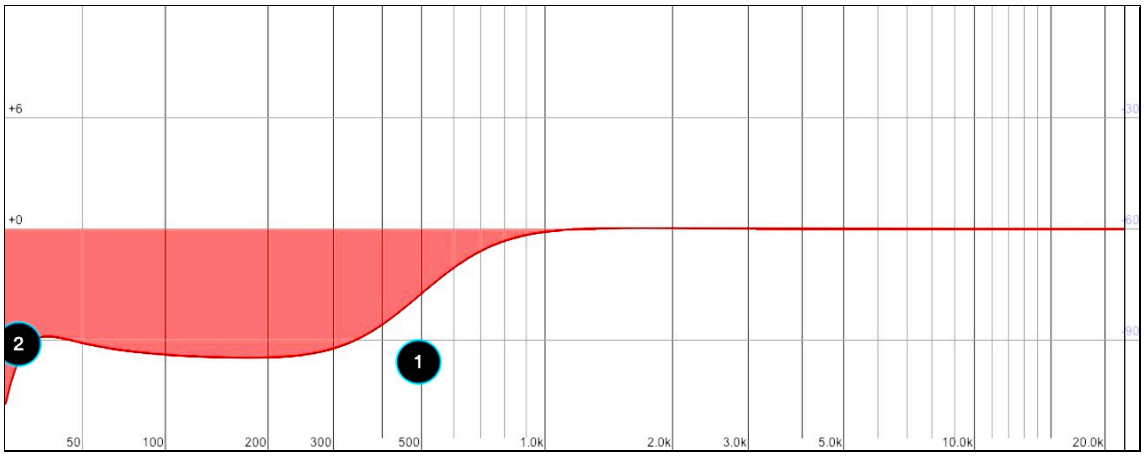



Figure 4.11 – EQ3: The fine tuning bass equalization filter for attenuating low frequencies on the headphones. Filter 1: Low Shelf, f1 = 500 Hz, Gain1 = -7 dB, Bandwidth = 0.8, Filter 2: High pass, f1 = 24.4 Hz, Gain1 = -2.3 dB, Bandwidth = 1.4.

| **Virtual Conditions** | **Bass Filter** |
|---|---|
| 7th Order | attenuated |
| 3rd Order | attenuated |
| 1st Order | attenuated |
| BRIR | increased |
| No Directivity | off |
| Short Reverb | attenuated |
| BRIR & No Directivity | increased |
| Lecture Room Acoustics | attenuated |
| Concert Hall Acoustics | attenuated |

Table 4.5 – With one exception, the different virtual acoustic conditions used one of two filters that affected low frequencies from the headphones. Compare figures 4.10 and 4.11.

# Chapter 5

# Method

First, this chapter gives a compact overview of the different conditions that were tested in the listening experiment. Within this overview, the four levels of "Degree of Freedom" are introduced. Furthermore, the experiment procedure and the response scale for the answers of the participants get explained. Before providing a list of all the different stimuli, this chapter presents the instructions that were given to the participants before they started the experiment.


Figure 5.1 – The mark for position 'A' in the real studio. This is the default position of the participants. Here they rated the stimuli that did not involve walking in the room.

All the stimuli in the listening experiment consisted of combinations of three different types of conditions: six different visual conditions, ten different acoustic conditions, and four different types of "degrees of freedom." Table 5.1 lists the names of all of these conditions. The visual and the acoustic conditions have been explained in detail in chapters 3 and 4, what is missing are the four types of DoFs, they are explained next.

| Visual Conditions | Acoustic Conditions | Degrees of Freedom |
|---|---|---|
| Dark Room | Real loudspeaker | without Rotation |
| Shoebox | 7th Order | with Rotation |
| Studio | 3rd Order | with Translation |
| Lecture Hall | 1st Order | with Interaction |
| Concert Hall | BRIR Reverb | |
| Reality | No Directivity | |
| | BRIR & No Directivity | |
| | Short Reverb | |
| | Lecture Room | |
| | Concert Hall | |

Table 5.1 – Overview of the different conditions in the three categories. All in all, 68 combinations out of these conditions were tested in the experiment. Table 5.3 provides a complete list of all those tested stimuli.

### Degrees of Freedom

The third category of conditions is the degree of freedom to which the participants were allowed to move their heads and bodies. The term "Degree of Freedom" is, in general, understood mathematically. The first three degrees stand for rotations around the three cartesian coordinate axes. The degrees of freedom four, five, and six correspond to translation along the coordinate axes. In this listening experiment, this mathematical system was extended with one more DoF, which has no mathematical representation: the interaction.

1. **Without Rotation**: Whenever the participants were asked to use this degree of freedom, they were not allowed to move their heads and their bodies while listening to the loudspeaker. They were asked to stand on the sign 'A' whenever this DoF was used.
2. **With Rotation**: The participants were asked to stand on the sign 'A', but here they were allowed to move their heads freely.
3. **With Translation**: The participants were allowed to move freely in the room and rotate their heads. They were not allowed to touch any of the walls or the interior.
4. **With Interaction**: The participants were allowed to move freely in the room and rotate their heads. Additionally, they were asked to touch the walls and the interior. One possibility was to sit on a desk in front of the loudspeaker. The desk was placed at the same location in the real studio and the virtual studio. The idea behind the DoF "with Interaction" was to investigate if the physical contact to the virtual environment influenced the plausibility of the presentation.

## 5.1 Experiment procedure

The listening experiment in this study was very comprehensive and, consequently, also took much time. Very fast participants finished in about 40 minutes, some needed 1.5 hours. There were 69 different stimuli in the experiment (compare table 5.3), and all of those were rated twice to be able to study the reliability of the individual participants.

The first thing the participants were asked to do was to read the instructions for the experiment. These instructions can be found in detail later in this chapter, and in german language in Appendix D. Before the real experiment started, the people did a short training. The purpose of this training was to familiarize the participants with the new environment. They saw the virtual rooms, they heard some of the acoustic models, and they tested the interface. Additionally, the training helped them with developing an "inner grid of plausibility." This grid was necessary because the study was not designed as an A/B-comparison experiment. Instead, the participants needed to compare their perceptions to their inner references for plausibility. This inner reference is an individual measure. Therefore, a definition of plausibility within the instructions tried to synchronize these inner references. Within the training, all the participants were presented with the same stimuli in the same order. y After the training, the real experiment started right away. The 138 stimuli were presented in an individual random order to every person in the experiment. Due to two visual conditions - the dark room and the real room - it was not possible to randomize the set of 138 stimuli completely. In the dark room, the participants had no orientation, so they would have been lost in the virtual room whenever they left the position 'A.' These dark conditions were only combined with the DoFs "without Rotation" and "with Rotation." Therefore, it was possible to present all of these stimuli at the beginning of the experiment. Within this block, they were randomized individually. The second exception was the visual reality. For all of the stimuli that included visual reality, it was necessary to take off the HMD. As this cannot be done randomly during the test - it would be far to effortful - all these stimuli were done at the end of the experiment in an individual random order. Within the experiment, two short breaks were scheduled, one after the first 50 stimuli and the second one after the second 50 stimuli. This gave the participants the chance to relax a little bit and to reflect on their perception. Many people stated after the listening experiment that it was quite tricky in the beginning, but especially after the first break, the rating became easier.

Within the experiment, the procedure looked like this: To start a stimulus, the participants had to press the "Play" button on their VR controller. With a rotation of the controller and the trigger button, they were able to choose a rating. Details on the scale and the mapping to the controller can be found in section 5.2. When they decided on a rating, the playback automatically stopped, and they saw their instructions for the next stimulus. Basically, these instructions informed the participants about the DoF for the upcoming stimulus. Whenever the next stimulus did not include translation or interaction, these instructions also told the people to walk back to position 'A' before pressing the "Play" button again.

## 5.2 Response Scale

| Ranks | Definitions |
|---|---|
| 7 | Fully plausible - "Acoustic authenticity is possible = it would be possible that a real loudspeaker would sound like this in such a room." |
| 4 | Neither-nor - "I realize that virtualization does not fit perfectly, but I don't find it completely implausible." |
| 1 | Fully implausible - "The sound and the picture do not fit together at all, or at least one of the above mentioned factors deviates completely from my expectations" |

Table 5.2 – This table provides the definitions of the seven Likert-type items in the response scale. The ratings one, four, and seven were defined, the space in between was left to the interpretation of the participants.

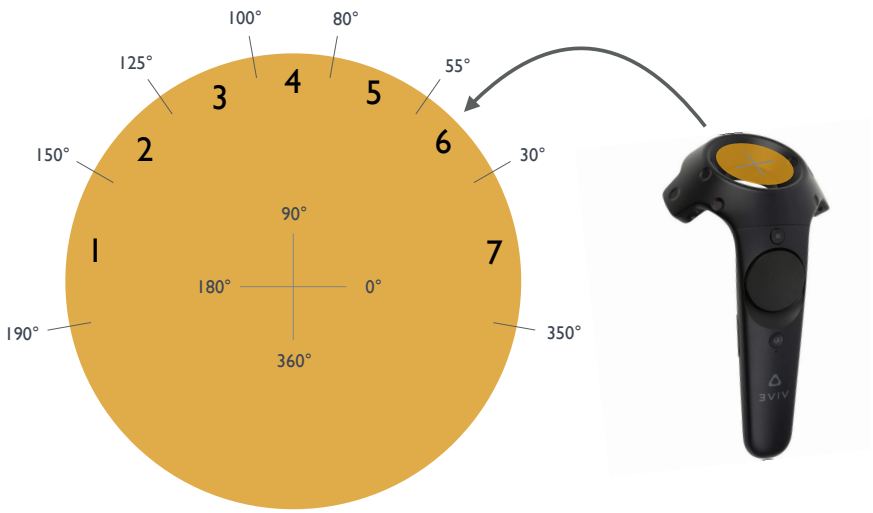



Figure 5.2 – The mapping of the Likert-style scale to the rotation of the controller.

To rate the plausibility of the different stimuli, the participants had a Likert-type scale with seven items [Lik32] at their disposal. Figure 5.2 shows how the scale was mapped to one of the VIVE controllers. The controller was turned around its main axis to choose between the different ratings. The numbers 1 to 7 were displayed via the HMD, see figure 2.5. When the participants decided for one of the ratings, they pressed the trigger button while holding the controller with the corresponding angle. Low ratings represented low plausibilities and high ratings represented high plausibilities. The mapping of the seven items was not done equally spaced because it is easier to hold the controller at 90° then at 0° and 180°. To give the participants a more comfortable feeling in the wrist, the extremes of the scale used 40° of the circle, the ratings 2, 3, 5, and 6 used 25°, and the central rating used only 20°. Table 5.2 shows the definitions of the ratings one, four, and seven. It was necessary to define the scale at these points to make sure all the participants use the same concept of plausibility.

## 5.3 Participant instructions

Before the listening experiment started, the participants were asked to read a few pages of instructions. These instructions explained everything they needed to know about the experiment. To be able to understand what the participants were told, the instructions will be explained here in detail. The original instructions in German language can be found in Appendix D. The participants were introduced to the test with these sentences: "In the following listening experiment, you will be asked to use virtual reality glasses and headphones to enter various virtual rooms. In these rooms, you will be able to see and to hear a virtual loudspeaker. The sound will be reproduced binaurally through your headphones. Your task in the virtual rooms will be to evaluate the plausibility of the presentations."

To define the concept of plausibility, the participants were told to think about these questions during the experiment:

- "Is this audiovisual presentation plausible?"
- "Does what I hear match what I see?"
- "Could a loudspeaker in such a room sound like this?"

Furthermore, they were told to pay attention, especially to the following parameters that were supposed to help them in the development of their "grid of plausibility."

- Externalization ... "Does this sound like the virtual loudspeaker is outside my head?
- Localization ... "Can I hear the virtual loudspeaker from the same direction I see it?"
- Distance ... "Do the acoustic and visual distances to the loudspeaker match? Does it sound closer or more distant than it looks?
- Timbre ... "Does the timbre of the loudspeaker change depending on your position in the room accordingly to your expectations ?" - Keep in mind that a real loudspeaker does not sound the same from all directions.
- Reverberation and room impression ... "Do your acoustic room impression and your visual room impression match? Does the room sound as big as it looks?"

The instructions also provided the definitions of the Likert-type items on the scale from 1 to 7, compare table 5.2. After the scale-definition, the participants were informed what happens when they decided on a rating: "When you have decided on a rating, your response is saved and the playback is paused. You will then see the instructions regarding your degree of freedom for the next stimulus."

There were four different instructions to prepare the participants for the next stimulus. When they prepared themselves, they pressed "Play" on the VR controller.

- Walk to point A | Do not move your head
- Walk to point A | Move your head
- Walk freely in the room | do not interact
- Walk freely in the room | interact

Last but not least the different conditions were explained like this:

A: The visual conditions:
   - A dark room: the VR goggles show a black picture.
   - One of four virtual environments.
   - The reality without VR goggles.

B: The acoustic conditions correspond to acoustic models of the loudspeaker in various qualities. Thus, it will be investigated which parameters of the acoustic virtualization are particularly crucial for a plausible representation.

C: The visual conditions:
   - Without rotation: You are standing the fixed position 'A' and should not move your head while listening. Look in the direction of the loudspeaker.
   - With rotation: You are again standing at the fixed position 'A', but you are allowed to turn your head.
   - With translation: You are allowed to move freely in the room, but should not have physical contact with the walls or the interior.
   - With interaction: You are allowed to move freely in space and should have physical contact with the room - for example, by leaning against a wall or sitting on a table.

## 5.4 Table with all tested stimuli

This section provides a table with a complete list of the 68 stimuli combinations that were used in the listening experiment. All of these stimuli were presented twice to be able to study the reliability of the participants. First, the participants were rating the stimuli involving the dark visuals, then all the virtual visuals, and then all the real visuals. Within these groups, the stimuli were sorted in a different random order for every participant. Within the random order, it was made sure that the repetitions of the stimuli with the numbers 58 to 62 did not occur close to each other. As these stimuli tend to be most different from the others, the participants otherwise might be able to remember their answers.

| # | VISUALS | ACOUSTICS | DOF |
|---|---|---|---|
| 1 | Dark | 1st Order | Without Rotation |
| 2 | Dark | 3rd Order | Without Rotation |
| 3 | Dark | 7th Order | Without Rotation |
| 4 | Dark | BRIR Reverb | Without Rotation |
| 5 | Dark | Real Loudspeaker | Without Rotation |
| 6 | Dark | No Directivity | Without Rotation |
| 7 | Dark | 1st Order | With Rotation |
| 8 | Dark | 3rd Order | With Rotation |
| 9 | Dark | 7th Order | With Rotation |
| 10 | Dark | BRIR Reverb | With Rotation |
| 11 | Dark | Real Loudspeaker | With Rotation |
| 12 | Dark | No Directivity | With Rotation |
| 13 | Shoebox | 7th Order | Without Rotation |
| 14 | Shoebox | Real Loudspeaker | Without Rotation |
| 15 | Shoebox | 7th Order | With Rotation |
| 16 | Shoebox | Real Loudspeaker | With Rotation |
| 17 | Shoebox | 7th Order | With Translation |
| 18 | Shoebox | Real Loudspeaker | With Translation |
| 19 | Studio | 1st Order | Without Rotation |
| 20 | Studio | 3rd Order | Without Rotation |
| 21 | Studio | 7th Order | Without Rotation |
| 22 | Studio | BRIR Reverb | Without Rotation |
| 23 | Studio | Real Loudspeaker | Without Rotation |
| 24 | Studio | Short Reverb | Without Rotation |
| 25 | Studio | No Directivity | Without Rotation |
| 26 | Studio | 1st Order | With Rotation |
| 27 | Studio | 3rd Order | With Rotation |
| 28 | Studio | 7th Order | With Rotation |
| 29 | Studio | BRIR Reverb | With Rotation |
| 30 | Studio | Real Loudspeaker | With Rotation |
| 31 | Studio | Short Reverb | With Rotation |
| 32 | Studio | No Directivity | With Rotation |

**Table 5.3 continued from previous page**

| # | VISUALS | ACOUSTICS | DOF |
|---|---|---|---|
| 33 | Studio | 1st Order | With Translation |
| 34 | Studio | 3rd Order | With Translation |
| 35 | Studio | 7th Order | With Translation |
| 36 | Studio | BRIR Reverb | With Translation |
| 37 | Studio | Real Loudspeaker | With Translation |
| 38 | Studio | Short Reverb | With Translation |
| 39 | Studio | No Directivity | With Translation |
| 40 | Studio | 7th Order | With Interaction |
| 41 | Studio | Real Loudspeaker | With Interaction |
| 42 | Lecture Room | 7th Order | Without Rotation |
| 43 | Lecture Room | Real Loudspeaker | Without Rotation |
| 44 | Lecture Room | 7th Order | With Rotation |
| 45 | Lecture Room | Real Loudspeaker | With Rotation |
| 46 | Lecture Room | 7th Order | With Translation |
| 47 | Lecture Room | Real Loudspeaker | With Translation |
| 48 | Lecture Room | 7th Order | With Interaction |
| 49 | Lecture Room | Real Loudspeaker | With Interaction |
| 50 | Concert Hall | 7th Order | Without Rotation |
| 51 | Concert Hall | Real Loudspeaker | Without Rotation |
| 52 | Concert Hall | 7th Order | With Rotation |
| 53 | Concert Hall | Real Loudspeaker | With Rotation |
| 54 | Concert Hall | 7th Order | With Translation |
| 55 | Concert Hall | Real Loudspeaker | With Translation |
| 56 | Concert Hall | 7th Order | With Interaction |
| 57 | Concert Hall | Real Loudspeaker | With Interaction |
| 58 | Studio | 7th Order Lecture Room | With Translation |
| 59 | Studio | 7th Order Concert Hall | With Translation |
| 60 | Studio | BRIR Reverb No Directivity | With Translation |
| 61 | Lecture Room | 7th Order Lecture Room | With Translation |
| 62 | Concert Hall | 7th Order Concert Hall | With Translation |
| 63 | Reality | 7th Order | Without Rotation |
| 64 | Reality | Real Loudspeaker | Without Rotation |
| 65 | Reality | 7th Order | With Rotation |
| 66 | Reality | Real Loudspeaker | With Rotation |
| 67 | Reality | 7th Order | With Translation |
| 68 | Reality | Real Loudspeaker | With Translation |

Table 5.3: All the stimuli that were used in the listening experiment.

# Chapter 6

# Results and Interpretations

In order to be able to analyze the data, different measures for the reliability of the participants are calculated and compared. This leads to the exclusion of one participant because of his or her poor reliability. Then the analysis strategy is presented, which uses a significance test and two measures for effect size.

## 6.1 Reliability of the Participants

In the experiment, all of the individual stimuli were presented twice to the participants. The reason for this is that now we can look for inconsistent behavior in the ratings to the same stimuli. If participants were not consistent in rating, it is likely they did not understand the question, were not concentrated or something unusual led to these results. Therefore after verification of such effects, it is possible to remove individual participants from the study.

The first technique to investigate the reliability of participants is to calculate a correlation coefficient. To do so, all the ratings of one participant are structured into two vectors. The two vectors contain the two ratings a participant gave to the same stimuli. If these vectors match perfectly, the correlation coefficient will become 1. Figure 6.1 shows the correlation coefficients for all 20 participants. The second way of testing is to investigate the mean individual deviations between the first and second answers of the participants. Those deviations were normed by the individual standard deviations of the data to compensate for different ranges of scale utilizations. Figure 6.2 shows these results.

Due to the results of participant number 11 in both investigations, his or her answers were removed from the dataset. Additionally to these results, participant number 11 was the only one rating the sound of the real loudspeaker with "neither-nor" while seeing reality and being able to walk in the real room. This could have been a mistake but is another indicator of his or her strange behavior in the experiment.

The correlation coefficients generally suggest that the participants had trouble reproducing themselves. On the one hand, this means they were not sure about their answers. On the other hand, this means their second answer might be independent of the first one. If so,

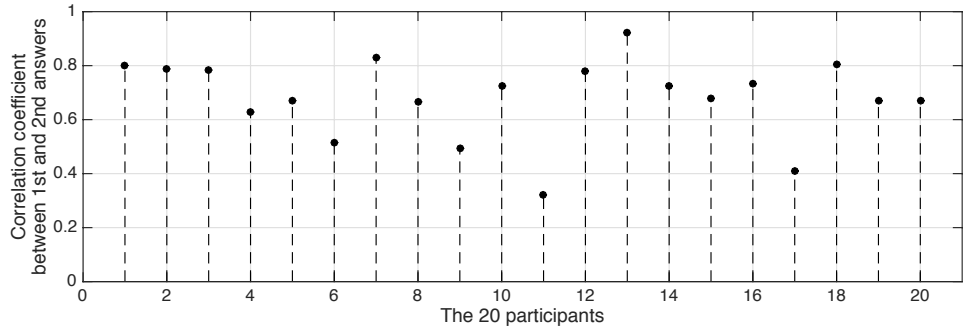



Figure 6.1 – The figure shows the correlation coefficients for the two answer vectors from every participant. Participant number 11 was the least consistent, resulting in a very low correlation coefficient of only $0.32$. The mean correlation coefficient is 0.68.

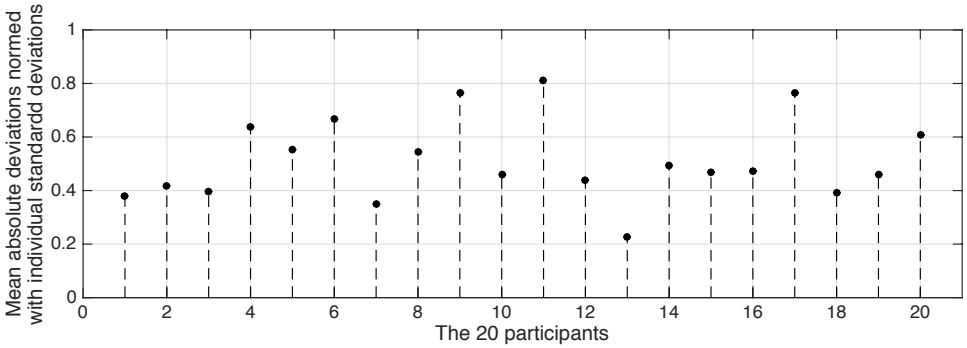



Figure 6.2 – The figure shows the mean absolute deviations between the individual answers of the 20 participants, normed with the individual standard deviations. Participant number 11 has the highest deviations.

it would be possible to think of a dataset with 40 individual ratings instead of 20. Therefore, the euclidean distances between the two answer-vectors of each participant were calculated (intra-rater distances). Also, the euclidean distances between the first answer-vector and the median first answer-vector of each participant was calculated (inter-rater distances, see figure 6.4).

Furthermore, the standard deviations for both of these difference-vectors (intra and intra differences) were calculated (see figure 6.3). With both measures, these intra-rater and inter-rater reliabilities do not differ much. Hence, for the rest of the analysis, the dataset was interpreted as if it were 38 independent ratings available per stimulus.

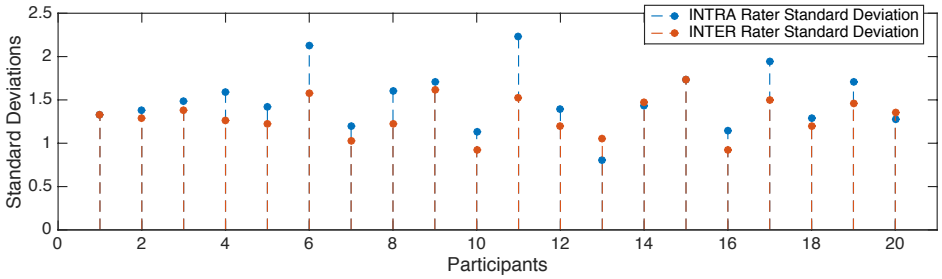



Figure 6.3 – The figure shows the intra-rater standard deviations and the inter-rater standard deviations.

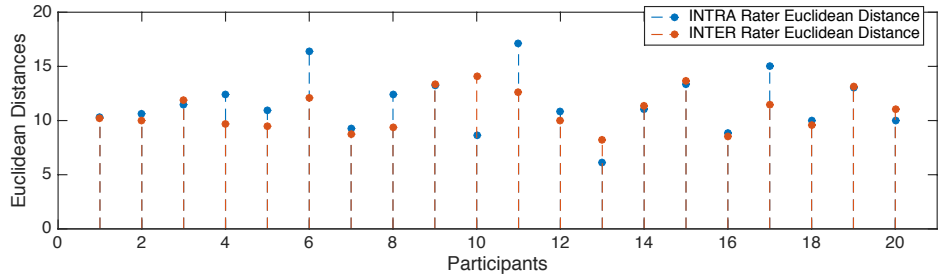



Figure 6.4 – The figure shows the intra-rater euclidean distances and the inter-rater euclidean distances. Note that the investigated vectors have 68 dimensions.

## 6.2 Analysis Strategy

Different data sets require different analysis strategies. The listening experiment in this study worked with at least an ordinal scale and therefore gained at least ordinal data. A common method of analysis for this type of data is the Wilcoxon Signed Rank Test [Wil45], which was used in combination with a Bonferroni-Holm Correction [Hol79]. This correction takes account of the multiple comparisons problem [RJ+12]. Furthermore, the effect size measures "Cliff's Delta" [Cli93] and "Cohen's d" [Coh13] are used to have a second indicator for true effects. Appendix A deals with the mathematics behind Cohen's d and Cliff's Delta. Note that Cliff's Delta and Cohen's d both can result in positive and negative values. In this thesis, we use the absolute values of both measures.

Whether it is possible to use Cohen's d for data which were collected with Likert-type items [Lik32] is something the literature does not agree on. Brown states it is sometimes possible to think of Likert scale data as interval data [Bro11]. As the scale in this study was defined for both ends (1 and 7) and in the center (4), there is a good chance that the gathered data is close to the interval level. Adams, Fagot and Robinson state in their book "A theory of appropriate statistics" [AFR65, p.100] that:

> *"Nothing is wrong per se in applying any statistical operation to measurements of given scale, but what may be wrong, depending on what is said about the results of these applications, is that the statement about them will not be empirically meaningful or else that it is not scientifically significant."*

| Effect Size | big | medium | small |
|---|---|---|---|
| Cohen's d | 0.80 | 0.50 | 0.20 |
| Cliff's Delta | 0.42 | 0.27 | 0.11 |

Table 6.1 – Cohen's d and Cliff's Delta compared to each other. See Appendix A.

In any case, to be able to use Cohen's d, the data needs to be normally distributed. The Jarque-Bera test indicates that most of the data is normally distributed ($\alpha = 5\%$). The test shows that only datasets that are very close to the borders of the rating scale are not normally distributed. As Cohen's d is a better-known measure than Cliff's Delta, it is used additionally to provide a better understanding of the effects in this study. Anyway, it is used only when the Jarque-Bera test states that the involved data sets follow a normal distribution and the data is are not bimodal. Table 6.2 provides the interpretation for Cliff's Delta in comparison to Cohen's d as calculated in Appendix A.

The p-values from the Wilcoxon-Signed Rank Test will be referred to as "p" from now on. Cliff's Delta and Cohen's d will be abbreviated with "$\Delta$" and "d".

## 6.3 Qualities of acoustic modeling

In the now following sections, the figures will be structured as follows: The ordinates will show the plausibilities from 1 (fully implausible) to 7 (fully plausible). The abscissas will mark the different acoustic stimuli that were presented. In most cases, these stimuli were presented with more than one DoF. These degrees can be identified with different markers in the figure. A legend explains which marker represents which DoF. Just for visualization, the figures show error bars with the mean values and their 95% confidence intervals. The data analysis always uses median values.

First, the different qualities of acoustic modeling are tested for their influence on plausibility. To do so, we look at the six different rooms and search for significant differences in the median values of the data. On the one hand, we look at the different acoustic models. On the other hand, we investigate how the different DoFs change the perception of these qualities. Only certain combinations of acoustic stimuli, visual stimuli, and DoFs were tested because to look at all possible combinations would have been far too much, and many combinations are not interesting.

### The 'Dark Room' model

Figure 6.5 shows the results of the listening experiment when the participants saw the 'Dark Room'. There were two DoFs and six acoustic conditions tested in this room. The two DoFs are 'without Rotation' and 'with Rotation'.

**Acoustic modeling Qualities**
Within the different acoustic conditions, we can see, that the *Real Loudspeaker* was al-

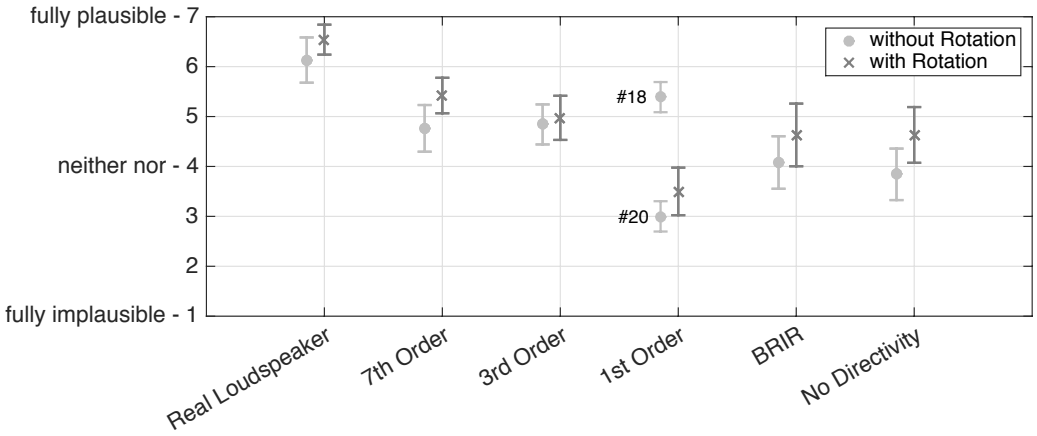



Figure 6.5 – Results in the 'Dark Room'. Mean values with 95% confidence intervals. *1st Order & with Rotation* resulted in a bimodal distribution. The sizes of the two modes are written next to the error bars.

ways rated with the highest plausibility. This holds for both DoFs ($p \leq 0.017$, $\Delta \geq 0.545$). We will see in the upcoming subsections that the *Real Loudspeaker* will be the most plausible for all the other visual conditions in this section. Without head rotation, the virtualizations only show one single significant difference between two of the stimuli. It is between *3rd Order* and *No Directivity* ($p = 0.0193$, $\Delta = 0.363$, $d = 0.721$). When the participants moved their heads the stimulus *No Directivity* became less plausible than *7th Order* ($p = 0.024$, $\Delta = 0.267$, $d = 0.562$) but more plausible than the *1st Order* ($p = 0.027$, $\Delta = 0.384$). In fact, with head rotation, *1st Order* becomes less plausible than *7th Order*, *3rd Order*, and *No Directivity* ($p \leq 0.0268$, $\Delta \geq 0.384$, $d \geq 0.727$). The only stimulus the *1st Order* is not significantly different from is *BRIR*; Cliff's Delta and Cohen's d both still detect medium to big effects ($p = 0.0652$, $\Delta \geq 0.372$, $d \geq 0.677$).

**Degrees of freedom**

Now we want to take a look at the influence of the Ambisonics order on the perception of the Ambisonics rotation. We investigate the development of the *7th*, *3rd*, and *1st Order* models concerning the DoFs. For *7th Order* a rotation improves the plausibility significantly ($p = 0.0102$, $\Delta = 0.273$, $d = 0.528$). With *3rd Order* the rotation is not changing anything ($p = 0.571$, $\Delta = 0.087$, $d = 0.104$) and with *1st Order* the rotation is downgrading the plausibility ($p = 0.045$, $\Delta = 0.236$).

**Interpretations**

The *No Directivity* model shows differences to the *7th Order* and the *3rd Order*. This seems unexpected as the participants were not allowed to move around and, therefore, should not have been able to hear a difference between *7th Order* and *No Directivity*. Anyway, without a directivity pattern, the loudspeaker activates the acoustics of the room in an omnidirectional way. This led to an investigation of the "Direct-to-Diffuse Ratio" of the different involved models. Table 6.2 shows that the DRR (see equation 6.1 with $h$ being an impulse response) for *No Directivity* is far less than the one for the other stimuli. According to [LILF08] the JND for DRRs between 0 and 10 dB lies between 2 and 4 dB.

| **Model** | 1st Order | 3rd Order | 7th Order | No Directivity |
|---|---|---|---|---|
| **DRR** | 5.6 dB | 6 dB | 5.5 dB | 1.3 dB |

Table 6.2 – Direct to Reverberation Ratios of the *1st Order*, *3rd Order*, *7th Order*, and *No Directivity* Models. No Directivity has a much lower ratio which explains why the participants were able to hear a difference between *7th Order* and *No Directivity* even without walking.

Therefore, the difference was audible and could have lead to an impression that was "to reverberant," even though the discrete reflections already had been attenuated by 4 dB.

Within the DoFs, another effect gets visible, which is known from the literature. It is the dependency of the binaural rendering quality of an Ambisonics rotation on the used Ambisonics order. In [ZFZ18] the authors suggest that the reason for this is the change in the perceived distance if a first-order Ambisonics signal is rotated and listened to binaurally. The distance error decreases with higher orders, which directly correlates to the perceived plausibility. One can also observe that the *1st Order* model is the only one that suffers from rotation. All other stimuli become more plausible when head movements are allowed. Also interesting within this context is that the head rotation even improved the plausibility of the *Real Loudspeaker* significantly, still, the effect size is small ($p = 0.008$, $\Delta = 0.138$). There are two possible explanations for this: First, without head rotation not all of the participants were sure if they heard the real loudspeaker, or second, the rotation itself improves the plausibility of acoustic reality (note, that also in real life people do not walk around without moving their heads).

$$DRR = 10log\frac{\int_0^T h^2(t)dt}{\int_T^\infty h^2(t)dt} \tag{6.1}$$

```
fs=44100;                                % Sampling rate
h = audioread('BRIR_r',[1 fs]);          % Load right channel of BRIR
E_direct = sum((h(1:170)).^2);           % Energy of direct sound
E_diffus = sum((h(171:end)).^2);         % Energy of diffuse sound
DRR = 10*log10(E_direct/E_diffus);       % Direct to Reverberant Ration
```

Listing 6.1 – Matlab code for the DRR calculated from a BRIR measurement.

## The 'Shoebox' model

The 'Shoebox' model is the one where the acoustic and visual modeling fit together best. This is due to the simple structure of the shoebox room without any interior. It consists only of four walls, a floor, and a ceiling that have exactly the same dimensions as in the acoustic model. In this virtual room, the two stimuli *Real Loudspeaker* and *7th Order* were studied.

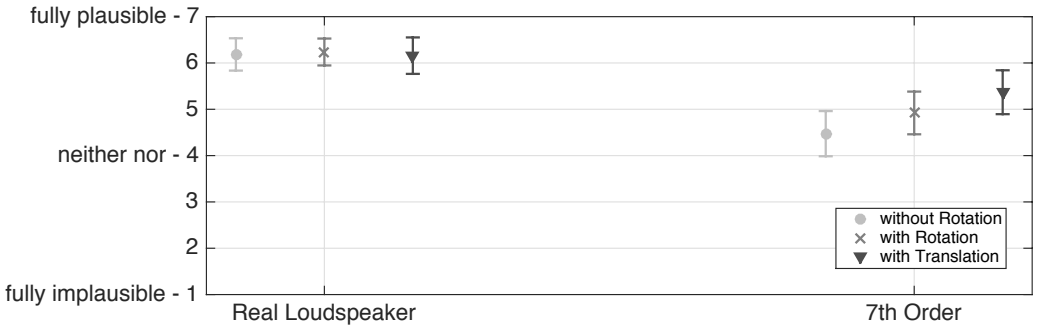



Figure 6.6 – Results in the 'Shoebox' model. Mean values with 95% confidence intervals.

**Acoustic modeling Qualities**
Regardless of the used DoF, the *Real Loudspeaker* was always rated more plausible then the *7th Order* ($p \leq 0.008$, $\Delta \geq 0.361$, $d \geq 0.604$).

**Degrees of Freedom**
If we have a look at the DoFs we see that the rating of the *Real Loudspeaker* is independent of them ($p \geq 1.937$, $\Delta \leq 0.026$, $d \leq 0.076$). The situation with the *7th Order* model is not the same. Here the DoFs have an influence that results in a significant difference between 'without Rotation' and 'with Translation' ($p = 0.014$, $\Delta = 0.347$, $d = 0.62$).

**Interpretations**
There are two possible interpretations of the fact that the *Real Loudspeaker* is independent of the DoF. The first one is that the *Real Loudspeaker* simply sounds so much more plausible than the virtualization that it is not essential what DoF is used. The other one is that the participants were able to identify the *Real Loudspeaker* and then rated it without further thinking about their answer. Compared to the *Real Loudspeaker*, it is visible that the DoFs influence the plausibility of the virtualization. Here the translation changes the rating significantly in comparison to the situation without head movement. Also, the higher the DoF is, the closer the virtualization comes to the *Real Loudspeaker*. It seems that plausibility is a property of the loudspeaker and the sound field itself. In the virtualization, the illusion becomes better with increasing DoFs, and thus the plausibility of the audiovisual presentation also improves.

## The 'Studio' model

The most detailed investigation was carried out in the 'Studio'. Here all ten acoustic stimuli and all four DoFs were part of the experiment. As in this subsection, we focus on the stimuli combinations which represent acoustic modeling qualities we will see only seven of the ten acoustic stimuli. It will also be the first time we will see the DoF 'Interaction'.

**Acoustic modeling qualities**
Again, we see the higher rating of the *Real Loudspeaker* in all DoFs ($p \leq 0.005$, $\Delta \geq 0.515$). The DoF 'without Rotation' features no significant differences between any of the virtualizations ($p \geq 0.752$, $\Delta \leq 0.161$, $d \leq 0.3$). 'With Rotation' features two differences:

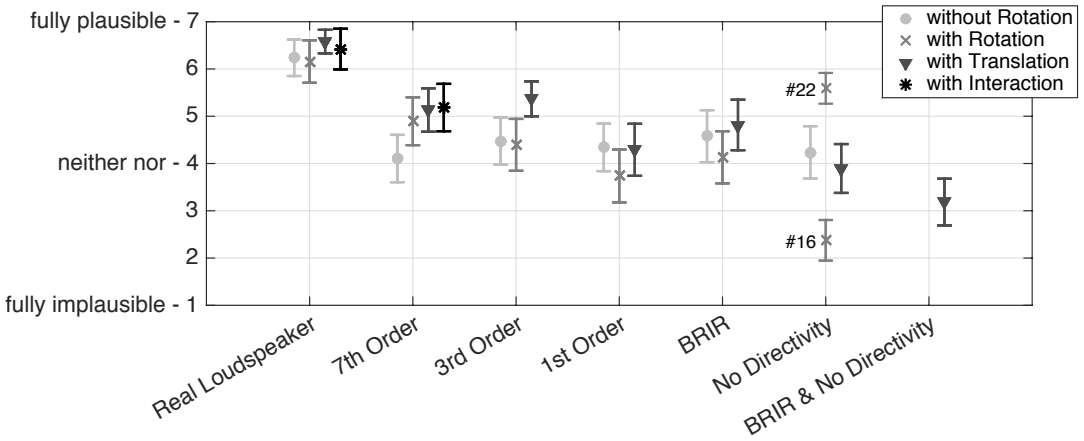



Figure 6.7 – Results in the 'Studio'. Mean values with 95% confidence intervals. *No Directivity* and 'with Rotation' resulted in a bimodal distribution. The sizes of the two modes are written next to the error bars.

| Stimuli | BRIR | No Directivity |
|---|---|---|
| **No Directivity** | 0.052 \| 0.324 \| 0.584 | |
| **BRIR & No Directivity** | 0 \| 0.538 \| 1.056 | 0.065 \| 0.249 \| 0.468 |

Table 6.3 – p | $\Delta$ | d - values for the comparisons between *BRIR*, *No Directivity*, and *BRIR & No Directivity* for the DoF 'with Translation'.

The first one is between the *7th Order* model and the *1st Order* model (p = 0.013, $\Delta$ = 0.375, d = 0.723). The second one is between the *7th Order* model and the *BRIR* model (p = 0.045, $\Delta$ = 0.253, d = 0.48), here the effects are medium. Furthermore, an interesting observation is that the two models *7th Order* and *3rd Order* are not significantly different for all DoFs ($p \geq 0.2032$, $\Delta \leq 0.167$, $d \leq 0.316$). The *No Directivity* stimulus is never significantly different from any other acoustic virtualization within the DoFs 'without Rotation' and 'with Rotation' ($p \geq 0.181$, $\Delta \leq 0.22$). Note that in this case the DoF 'with Rotation' led to a bimodal distribution, therefore it is not possible to calculate Cohen's d for these comparisons. Only when the participants were allowed to move freely in the room, the directivity pattern of the loudspeaker had an influence on the plausibility. This lead to significant differences to the *7th* and *3rd Order* model ($p \leq 0.013$, $\Delta \geq 0.44$, $d \geq 0.842$). Within the DoF 'with Translation' an additional acoustic quality was tested: *BRIR & No Directivity*. This stimulus essentially is the combination out of *BRIR* and *No Directivity* and also the plausibility ratings suggest this - see table 6.3. If the Bonferroni-Holm correction was a little less conservative all of these comparisons were at least weakly significant.

**Degrees of Freedom**

Concerning the dependency on the DoFs, one can observe the same tendency as in the 'Dark Room': The quality of Ambisonics rotation is dependent on the order of the Ambisonics signal. With *7th Order* the rotation increases the plausibility compared to the *No Rotation* stimulus (p = 0.033, $\Delta$ = 0.285, d= 0.523), with third order the rotation does not change the plausibility (p = 0.856, $\Delta$ = 0.003, d = 0.05), and with first order figure

6.7 suggests that plausibility gets worse. Anyway, this is not a significant difference ($p = 0.27$, $\Delta = 0.2$, $d = 0.379$).

Again, we see the independence of the *Real Loudspeaker* from the DoFs ($p \geq 0.154$, $\Delta \leq 0.195$ $d \leq 0.386$). The DoF 'Interaction' does not change plausibility in comparison to 'Translation', neither for the *Real Loudspeaker* ($p = 1.073$, $\Delta = 0.028$, $d = 0.15$), nor for the *7th Order* model) ($p = 0.86$, $\Delta = 0.035$, $d = 0.036$).

**Interpretations**
The DoF 'Interaction' is not increasing the plausibility. Here the investigated question was if physical contact to the virtual environment improves the illusion and if that also led to an improvement of the plausibility of the acoustic representation. It seems like this is not the case. What could be true is that a short moment of distraction might lead to a higher acceptance of the presented virtual acoustics. It also seems to be true that the plausibility of acoustic virtualizations is time-dependent. Staying in VR longer could lead to a better acceptance of the new virtual acoustics, which would lead to higher plausibility.

Furthermore, the ratings for all of the acoustic models most likely have been distorted, due to the prominent presence of the *Real Loudspeaker*. Many participants were able to identify the *Real Loudspeaker* and rated it with high plausibility, even though the participants were instructed to rate every stimulus independently.

Figure 6.7 shows that the DoF 'Translation' helped the plausibility in all cases but the one without loudspeaker directivity. This makes sense as the translation is beneficial in hearing an effect of the directivity pattern in the direct sound.

The observation that the *7th Order* model and the *3rd Order* model are never significantly different from each other truly affects the practice in virtual reality audio production. It results in the possibility to reach the same level of acoustic plausibility with 16 Ambisonics channels compared to 64 channels. First-order modeling results in a significant loss of plausibility unless the application is working without head movement, which is not likely in modern VR environments.

The decrease of plausibility over the stimuli *BRIR*, *No Directivity*, and *BRIR & No Dorectivity* for the DoF 'with Translation' supports the hypothesis that a combination of rendering quality reductions leads to a reduction of plausibility.

## The Reality

**Acoustic modeling Qualities**
Just like in the virtual rooms, also here the *Real Loudspeaker* always was rated significantly better than the *7th Order* model, regardless of the DoFs ($p < 0.001$, $\Delta \geq 0.618$).

**Degrees of Freedom**
Again, we see no significant dependencies on the DoFs for the *Real Loudspeaker* ($p \geq 0.082$, $\Delta \leq 0.19$). Just like in the other virtual rooms the *7th Order* model is dependent on the DoFs. There is no significant difference between 'without Rotation' and 'with Rotation' ($p = 0.127$, $\Delta = 0.152$, $d = 0.3056$), but both of them are different from 'with Translation' ($p \leq 0.001$, $\Delta \geq 0.221$, $d \geq 0.511$).


Figure 6.8 – Results in the real studio. Mean values with 95% confidence intervals.

**Interpretations**
The fact that some participants rated the *Real Loudspeaker* not fully plausible while seeing the reality without VR googles is especially interesting. It seems like the simple fact that the participants were wearing headphones influenced their perception, and some of them were not 100% sure it was the *Real Loudspeaker* that was playing. Otherwise, there is no reason to rate reality itself not fully plausible, hence authentic. Furthermore, one can again observe that the *Real Loudspeaker* is independent of the DoFs while the virtualization is dependent on it. Just like in the shoebox model, the virtualization gains plausibility with increasing DoF. Still, it is not enough to reach the level of the *Real Loudspeaker*.

## 6.4 Virtual Room Divergence Effect

Inspired by the research of Stephan Werner [WKMB16], this part of the experiment investigated a possible "Virtual Room Divergence Effect." This would result in a lower rating of plausibility whenever the participants are listening to a small room while seeing a big room via their HMD. It is necessary to state that this must not be confused with the question Werner Stephan asked his participants in [WKMB16]. There, the participants rated the perceived externalization. In this study, the participants were asked to rate plausibility, which itself was, by definition, dependent on externalization. Therefore a possible "Virtual Room Divergence Effect" (VRDE) is defined here to be a "dependency of the plausibility of an acoustic stimulus on the visually perceived size of the surrounding virtual room."

### In the 'Studio'

In the 'Studio', three different acoustic models were used to study the VRDE. These stimuli were *Short Reverb*, *Lecture Room* and *Concert Hall*. All of these rooms did not have the same acoustic properties as the studio room. Figure 6.9 shows the results of these three acoustic models in the 'Studio'. All of them result in very low plausibility ratings. An investigation of the DoFs for the *Short Reverb* results in no significant differences ($p \leq 0.171$, $\Delta \leq 0.224$, $d \leq 0.415$).

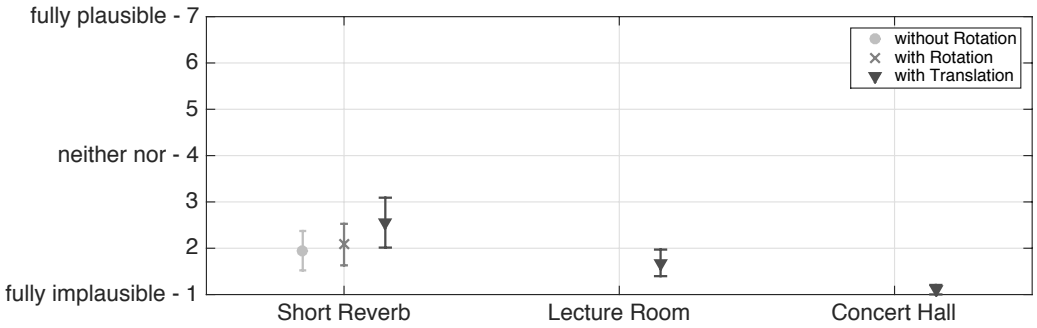



Figure 6.9 – Results in the 'Studio' with the acoustic stimuli from the category "Virtual Room Divergence Effect". Mean values with 95% confidence intervals.

**Interpretation**
Both, a too short reverb (*Short Reverb*) and a too long reverb (*Lecture room & Concert Hall*) reduce the plausibility. In the first case, the shorter reverb leads to less externalization, and therefore the plausibility suffers. Furthermore, this acoustic model sounds too dry for a room like the 'Studio' to sound like this in reality. In the second case, the small studio room simply cannot have a reverberation time like a lecture room or a concert hall. Under certain circumstances, big rooms can have short reverb times, but small rooms usually do not sound very big.

## In the 'Lecture Room'

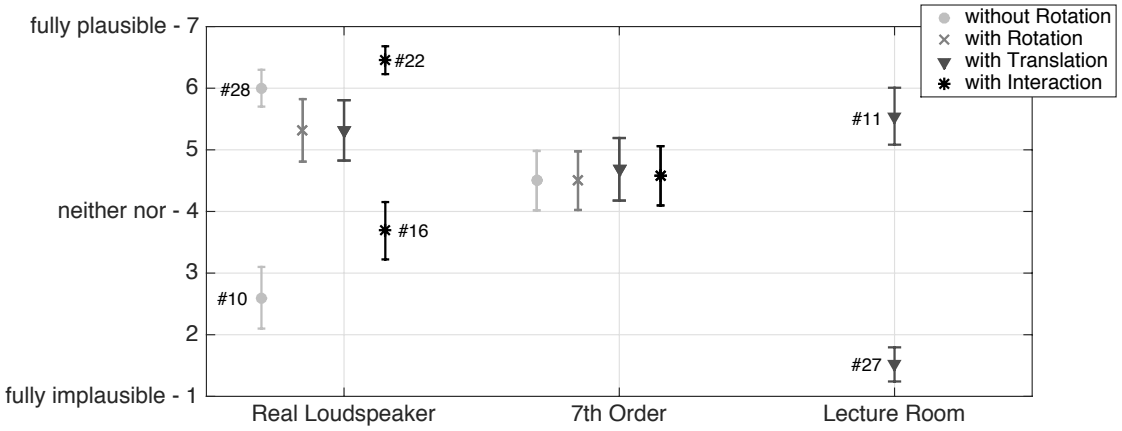
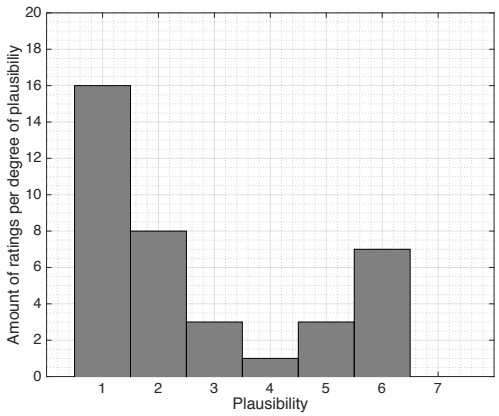


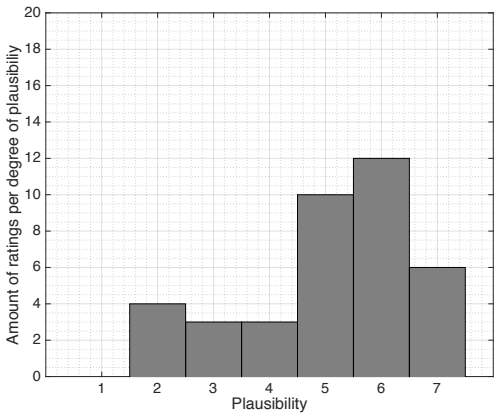

Figure 6.10 – Results in the 'Lecture Room' with the *Real Loudspeaker* from the studio, the *7th Order* model from the studio and an the acoustic *Lecture Room*. For some of the stimuli, the participants split up into two groups. In these cases, the data was clustered and two datasets are plotted. The number of answers per cluster is plotted next to the error bar. The error bars show mean values with 95% confidence intervals.

As can be seen in figure 6.10, the acoustic model of the 'Lecture Room' has not lived up to the expectations of most participants. The histograms in figure 6.11 show that the

participants were divided into two groups: The ones who thought the acoustic and the visual model did not fit together, and the ones who thought they did. Why is this? Figure 6.12 shows that the people who rated the acoustic *Lecture Room* low were likely to be the same people that rated the *Real Loudspeaker* high. A plausible theory here is that most of the participants did decide very fast without thinking about their decision. As this acoustic model is part of the category "Visual Room Divergence Effect," it was not featured as often as the several acoustic studio models. Therefore, chances were high that the participants were surprised when they heard this acoustic stimulus and intuitively choose a bad rating simply because they were not used to this sound. The real lecture room also has quite a long reverb, which is unexpected for its size. Due to this circumstance, another possible explanation for the bimodality is that only a few of the participants were familiar with the acoustics of the real room and therefore had the chance to rate the model as rather plausible. The ratings of the *Real Loudspeaker* and the *7th Order* model are also interesting. The *Real Loudspeaker* includes two bimodal distributions, one for the DoF 'without Rotation' and one for 'with Interaction'. Still, the Wilcoxon Signed Rank test and Cliff's Delta state that the rating of the *Real Loudspeaker* was not dependent on the DoFs ($p \geq 1.819$, $\Delta \leq 0.069$). Furthermore, the *7th Order* model for the first time does not show a dependency on the DoFs in this room ($p \geq 1.671$, $\Delta \leq 0.063$, $p \leq 0.125$). It seems that for their rating, the participants focused on something that was not changing with respect to their position or head rotation. Most likely, this was the coloration of the sound and the reverb in the room.




(a) A Histogram with the data in the 'Lecture Room'. There are two groups within the participants. The ones who think the acoustic model is not plausible and the ones who think it is.




(b) A Histogram with the data in the 'Concert Hall'. Here, the participants agreed more and rated the situation as rather plausible.

Figure 6.11 – Histograms showing the number of ratings per degree of plausibility for (a) the 'Lecture Room' and (b) the 'Concert Hall'. In both these situations, the participants were presented with the acoustic and visual models of the same rooms.

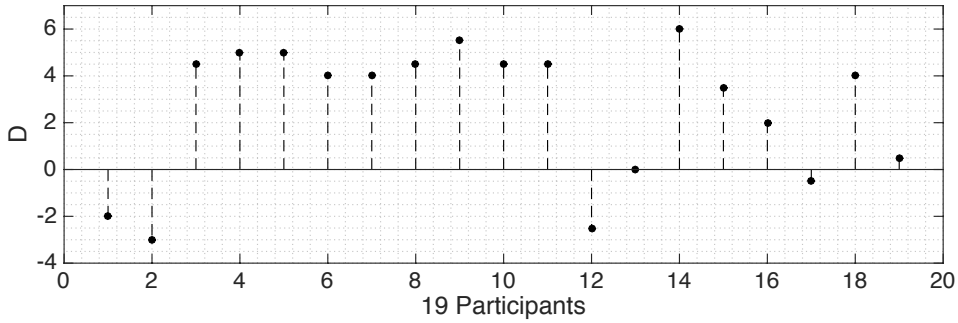



Figure 6.12 – D = Differences between the average ratings for the *Real Loudspeaker* with interaction and the acoustic lecture room for each participant. High ratings belong to participants that rated the *Real Loudspeaker* high and the virtual one low. Low values represent participants that rated the *Real Loudspeaker* low and the virtual one high.

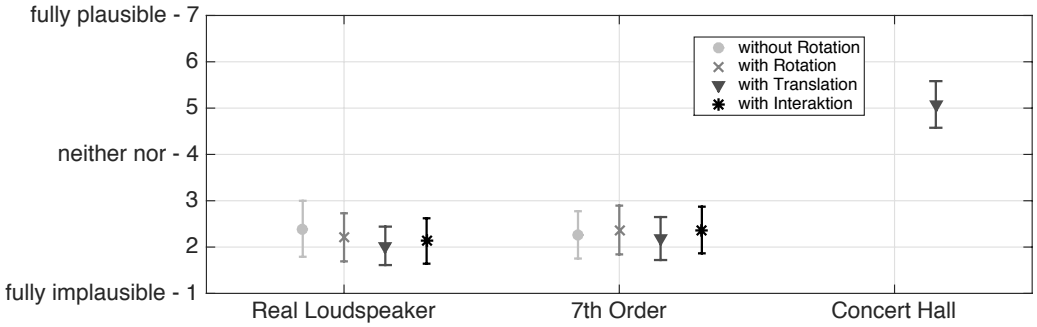



Figure 6.13 – Results in the 'Concert Hall'. Mean values with 95% confidence intervals.

## In the 'Concert Hall'

In the concert hall also the concert hall acoustics were rated most plausible. Just like in the 'Lecture Room' the *Real Loudspeaker* and the *7th Order* model are independent of the DoFs (*Real Loudspeaker*: $p \geq 0.589$, $\Delta \leq 0.009$, $p \leq 0.237$; *7th Order*: $p \geq 1.578$, $\Delta \leq 0.047$, $p \leq 0.127$). Here it seems to be the very short reverb of the studio acoustics that is implausible for such a big room. Figure 6.11(b) shows that unlike in the 'Lecture Room' here the participants were not separated into two distinct groups.

Concerning both bigger rooms, the lecture room, and the concert hall, one can conclude: In the first one, the room does not look big enough for the acoustics of the studio to become fully implausible. The concert hall simply cannot sound like the studio - here these acoustics become implausible.

## 6.5 Pooled data analysis

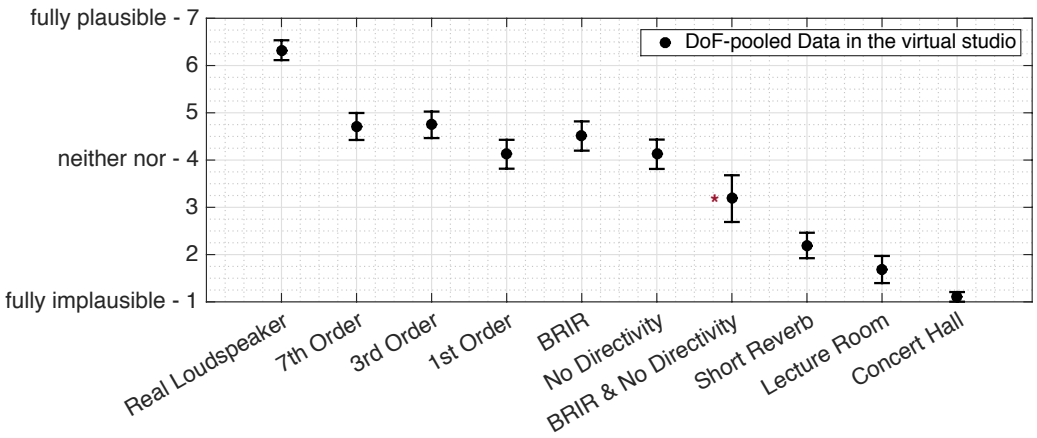



Figure 6.14 – Results in the 'Studio' with pooled DoFs. Mean values with 95% confidence intervals. The error bars consist of the DoFs 'without Rotation', 'with Rotation', and 'with Translation', except for *BRIR & No Directivity*. Here, only 'with Translation' was tested, it is marked with a red star next to the error bar.

This section provides some of the data pooled over the degrees of freedom. This way it is possible to investigate the impact of only the acoustic modelings and the visuals that were presented to the participants.

The first analysis in this context is done with data from the 'Studio', see figure 6.14. The 'Studio' was the only room where all the acoustic models from the experiment were tested, see table 5.3 lines 19-41 and 58-60. After pooling the DoFs, the ratings depend only on the acoustic models. One can see that also here the *Real Loudspeaker* model is always significantly better than all the binaural models ($p < 0.001$, $\Delta \geq 0.61$, $d \geq 1.188$). Especially interesting is that the *7th Order* model is not significantly different from both the *3rd Order* model ($p = 1.723$, $\Delta = 0.015$, $d = 0.023$) and from the *BRIR* model ($p = 0.584$, $\Delta = 0.062$, $d = 0.0126$). It seems that dynamic reverberation modeling is not as important as dynamic direct sound modeling. If the direct sound arrives at the ears with the 7th or 3rd-order signal, then it is sufficient to play back the reverberation via a static BRIR-reverb convolution. If the acoustic modeling combines the BRIR model with a direct sound that has no directivity pattern, then this results in reduced plausibility. The *1st Order* is not significantly different from the *No Directivity* model ($p = 1.723$, $\Delta = 0.004$, $d \ll 0.001$). Comparing *Short Reverb*, *Lecture Room*, and *Concert Hall* to each other, one finds that they also significantly differ from each other ($p \leq 0.014$, $\Delta \geq 0.043$, $d \geq 0.427$).

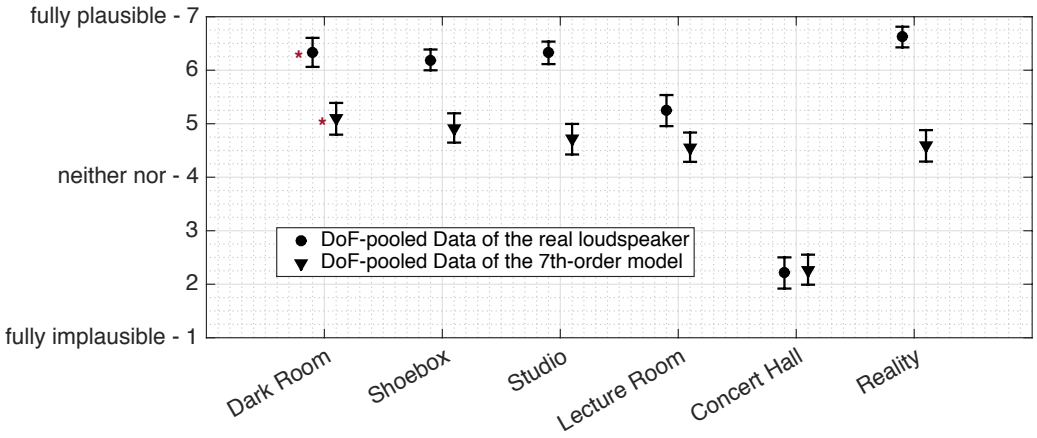



Figure 6.15 – Results of the *Real Loudspeaker* and the *7th order* model in all the rooms with pooled DoFs. Mean values with 95% confidence intervals. The error bars consist of the DoFs 'without Rotation', 'with Rotation', and 'with Translation', except for the stimulus *Dark Room*. Here, only 'with Rotation' and 'without Rotation' were tested, they are marked with a red star next to the error bars.

Also interesting is the influence of the visuals on the perception of the plausibility. To study this influence, the data of the *Real Loudspeaker* and the *7th Order* were pooled with respect to the DoFs and plotted for each of the six rooms in figure 6.15. The *Real Loudspeaker* was never significantly different between the visuals 'Dark Room', 'Shoebox', 'Studio, and 'Reality' ($p \geq 0.56$, $\Delta \leq 0.072$, $d \leq 0.267$). The 'Lecture Room' and the 'Concert Hall' show lower ratings. When comparing the 'Studio' with the 'Lecture Room' ($p \ll 0.001$, $\Delta = 0.442$, $d = 0.793$) and the 'Concert Hall' ($p \ll 0.001$, $\Delta = 0.93$, $d = 3.017$), one can observe that there is a degrading of plausibility. This seems to be dependent on the size of the displayed room. Future research could investigate if the plausibility of an acoustic model correlates to the size of the virtual room. Studying the *7th Order* in figure 6.15, one can see that here, compared to the 'Studio', the 'Lecture Room' was not degrading the plausibility significantly ($p = 0.73$, $\Delta = 0.065$, $d = 0.099$). For the *7th Order* only the 'Concert Hall' lead to a significant difference when compared to all other visuals ($p \ll 0.001$, $\Delta \geq 0.532$, $d \geq 1.539$). It seems that the *Real Loudspeaker* has some inherent quality that the virtual one is missing, which leads to an improved capability to estimate the size of the room.

## Feedback by the participants

Some of the feedback that was given by the participants is listed below:

- In the dark room, none of the acoustic stimuli were fully implausible.
- Immediately after listening to a stimulus that had a lot of reverb, I tended to rate the next one less plausible. It felt less externalized when the reverb was suddenly gone. Also, after the real loudspeaker, the plausibility was lower.
- The virtual loudspeaker sound very good if my position is about 60 degrees of the central axis. In front of the loudspeaker, the modeling is less plausible.
- The virtual loudspeaker sounded less plausible if I went closer to it.
- The directivity pattern of the loudspeaker was very important for my rating.
- After the first break in the experiment, it was much easier to rate the plausibility.
- The virtual loudspeaker's bass frequencies were missing in the acoustic concert hall model.
- In the shoebox model, the virtual loudspeaker sounded best.

# Chapter 7

# Conclusion and Outlook

This chapter recapitulates the found results and, at the same time, reconsiders their significance for modern applications of virtual reality. It also provides ideas for future research.

The goal of this thesis was to contribute to the research in virtual acoustics for applications in virtual reality. At the beginning of this thesis, the word-pairs "immersion" & "presence" and "plausibility" & "authenticity" have been established. The listening study investigated the second pair. In order to conduct such an experiment, it was necessary to implement a flexible test environment first. The experiment then asked 20 participants to rate the plausibilities of 68 different stimuli-combinations. Maximal plausibility was defined to be "authentic." Since authenticity is defined to be indistinguishable from reality [Pel01], also reality was presented in the form of a real room and the acoustics of a real loudspeaker.

## Conclusion

We already knew that head rotations improve the plausibility of binaural rendering in general. This is true for VR situations too, but that is not everything. In the conducted experiment, the participants were also able to walk through the virtual rooms, which improved the plausibility of the audio-visual presentation even further. What did not improve the ratings in comparison to the DoF 'Translation' was the 'Interaction'. Even though this condition stimulated the tactile sense additionally to the vision and the auditory system, the plausibility was not increasing.

The *7th Order* model and the *3rd Order* model were not significantly different from each other, and we observed only very small effect sizes. This means that the quality of a 7th-order rendering can also be reached with 3rd order. This results in a reduction of the number of necessary Ambisonics channels from 64 to 16. Subsequently, this circumstance makes it easier to efficiently implement binaural rendering in game engines and for Ambisonics streaming.

The results related to the BRIR reverb suggest that an efficient implementation could also be done with a static reverberation. As long as the direct sound was modeled dynamically and with a directivity pattern, the participants rated this stimulus also with high ratings.

In general, a directivity pattern improved the plausibility of the virtual loudspeaker, especially when the participants were allowed to move in the room.

The stimulus with unusually short reverberation resulted in all cases in a low plausibility rating. There are two reasons for that: the first one is that reverb plays an essential role in the externalization of binaural audio [GWH19]. Without proper externalization, the users hear the sound in their heads, and this results in low plausibility. The second reason is that the "Virtual Room Divergence Effect" becomes relevant. This effect is the consequence of a disproportion between the user's visual and acoustic impression. The room looks bigger than it sounds, and this leads to a low plausibility rating. This effect can be observed for big rooms with short reverb and even more for small rooms with long reverb (as this is less likely in reality).

The virtual studio and the virtual shoebox model did not result in different ratings. The assumption is reasonable that the visual modeling of interior, doors, and windows is not improving plausibility in comparison to an empty room of the same size.

It is important to remember that this listening experiment investigated plausibility under the most challenging circumstances, hence with reality as a repetitive stimulus in the test. On the one hand, this made sure the top of the scale indeed was authenticity. On the other hand, it is reasonable that the sheer presence of the real loudspeaker influenced the other ratings by becoming a reference. When the loudspeaker would not have been there, we most likely would have seen higher ratings for some virtualizations. Anyway, we would not have known their relation to authenticity.

Another important consideration is that plausibility is not only a question of how the stimuli are presented but also how they are received. A person using a virtual reality application usually is not actively thinking about their acoustic environment. It is instead the case that they concentrate on something completely different and will perceive the virtual acoustics subconsciously. In such a situation, a plausible impression is far easier to achieve. Another factor might be the time the users are exposed to the virtual acoustics without being remembered of the quality of real acoustics. It seems that the brain can accept the virtual acoustic over time, which would result in a rising plausibility.

# Outlook

The possible time dependency of the acceptance of virtual acoustics is most certainly a topic for future investigation. This would be possible for example by asking participants the same questions more often with them playing any virtual game in between. Furthermore, it might also be interesting to investigate the virtual room divergence effect in more detail. It would be possible to let the participants change the size of the room until they think it fits the acoustic presentation.

Generally, one has to mention that the experiment conducted in this project was very comprehensive. For detailed and more precise data, one would have to focus on specific aspects and do fewer combinations.

# Bibliography


[AFR65] E. W. Adams, R. F. Fagot, and R. E. Robinson, "A theory of appropriate statistics," *Psychometrika*, vol. 30, no. 2, pp. 99–127, 1965.

[Blo04] P. Bloomfield, *Fourier analysis of time series: an introduction*. John Wiley & Sons, 2004.

[BMJ20] S. Bin, S. Masood, and Y. Jung, "Virtual and augmented reality in medicine," in *Biomedical Information Technology*. Elsevier, 2020, pp. 673–686.

[Bro11] J. D. Brown, "Likert items and scales of measurement?" *Shiken: JALT Testing & Evaluation SIG Newsletter*, pp. 10–14, March 2011.

[BV17] L. P. Berg and J. M. Vance, "Industry use of virtual reality in product design and manufacturing: a survey," *Virtual reality*, vol. 21, no. 1, pp. 1–17, 2017.

[Cli93] N. Cliff, "Dominance statistics: Ordinal analyses to answer ordinal questions." *Psychological bulletin*, vol. 114, no. 3, p. 494, 1993.

[CMM05] S. Carlile, R. Martin, and K. McANALLY, "Spectral information in sound localization," *International review of neurobiology*, vol. 70, pp. 399–434, 2005.

[Coh13] J. Cohen, "Statistical power analysis for the behavioral sciences," May 2013. [Online]. Available: http://dx.doi.org/10.4324/9780203771587

[EFH20] K. Enge, M. Frank, and R. Höldrich, "Listening experiment on the plausibility of acoustic modeling in virtual reality," *Fortschritte der Akustik, DAGA*, 2020.

[FB19] M. Frank and M. Brandner, "Perceptual evaluation of spatial resolution in directivity patterns," *Fortschritte der Akustik, DAGA, Rostock*, 2019.

[FRB20] M. Frank, D. Rudrich, and M. Brandner, "Augmented practice-room - augmented acoustics in musikc education," *DAGA*, 2020.

[Ger73] M. A. Gerzon, "Periphony: With-height sound reproduction," *Journal of the audio engineering society*, vol. 21, no. 1, pp. 2–10, 1973.

[GWH19] P. M. Giller, F. Wendt, and R. Höldrich, "The influence of different brir modification techniques on externalization and sound quality," 2019.

[Hof16] M. Hofer, *Presence und involvement*. Nomos Verlagsgesellschaft mbH & Co. KG, 2016.

[Hol79] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979. [Online]. Available: http://www.jstor.org/stable/4615733

[JEK19] D. Johnston, H. Egermann, and G. Kearney, "Measuring the behavioral response to spatial audio within a multi-modal virtual reality environment in children with autism spectrum disorder," *Applied Sciences*, vol. 9, no. 15, p. 3152, 2019.

[KA03] L. Krabbendam and A. Aleman, "Cognitive rehabilitation in schizophrenia: a quantitative analysis of controlled studies," *Psychopharmacology*, vol. 169, no. 3-4, pp. 376–382, 2003.

[Kas20] D. Kasprowicz, "Virtual embodiment," in *Handbuch Virtualität*. Springer, 2020, pp. 385–402.

[Kel16] P. Kellnhofer, "Perceptual modeling for stereoscopic 3d," Ph.D. dissertation, Saarland University, 2016.

[KR11] C. Kuhn-Rahloff, "Prozesse der Plausibilitätsbeurteilung am Beispiel ausgewählter elektroakustischer Wiedergabesituationen," Ph.D. dissertation, Technische Universität Berlin, 2011.

[Lik32] R. Likert, "A technique for the measurement of attitudes." *Archives of psychology*, 1932.

[LILF08] E. Larsen, N. Iyer, C. R. Lansing, and A. S. Feng, "On the minimum audible difference in direct-to-reverberant energy ratio," *The Journal of the Acoustical Society of America*, vol. 124, no. 1, pp. 450–461, 2008. [Online]. Available: https://doi.org/10.1121/1.2936368

[LR] A. Lindau and S. Roos, "Perceptual evaluation of discretization and interpolation for motion-tracked binaural (mtb) recordings (perzeptive evaluation von diskretisierungs-und interpolationsansätzen."

[MA20] Z. H. Majeed and H. A. Ali, "A review of augmented reality in educational applications," *International Journal of Advanced Technology and Engineering Exploration*, vol. 7, no. 62, pp. 20–27, 2020.

[MBWS06] U. Mögerle, S. Böcking, W. Wirth, and H. Schramm, "Unterhaltungserleben in virtuellen Medien. Die Rolle von Medieneigenschaften und Personenmerkmalen beim Entstehen von Spatial Presence," *Empirische Unterhaltungsforschung: Studien zu Rezeption und Wirkung von medialer Unterhaltung*, vol. 8, p. 87, 2006.

[MKRB+20] N. Meyer-Kahlen, D. Rudrich, M. Brandner, S. Wirler, S. Windtner, and M. Frank, "DIY Modifications for Acoustically Transparent Headphones," in *AES 148th Convention, e-Brief 61*, 2020.

[NHBH20] M. Noghabaei, A. Heydarian, V. Balali, and K. Han, "Trend analysis on adoption of virtual and augmented reality in the architecture, engineering, and construction industry," *Data*, vol. 5, no. 1, p. 26, 2020.

[Pel01] R. S. Pellegrini, "Quality assessment of auditory virtual environments." Georgia Institute of Technology, 2001.

[PH07] D. Pinchon and P. E. Hoggan, "Rotation matrices for real spherical harmonics: general rotations of atomic orbitals in space-fixed axes," *Journal of Physics A: Mathematical and Theoretical*, vol. 40, no. 7, p. 1597, 2007.

[PS19] V. Pulkki and U. P. Svensson, "Machine-learning-based estimation and rendering of scattering in virtual reality," *The Journal of the Acoustical Society of America*, vol. 145, no. 4, pp. 2664–2676, 2019.

[Ray07] L. Rayleigh, "On our perception of sound direction," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 13, no. 74, pp. 214–232, 1907. [Online]. Available: https://doi.org/10.1080/14786440709463595

[RJ+12] G. Rupert Jr *et al.*, *Simultaneous statistical inference*. Springer Science & Business Media, 2012.

[RKCS06] J. Romano, J. D. Kromrey, J. Coraggio, and J. Skowronek, "Appropriate statistics for ordinal level data: Should we really be using t-test and cohen'sd for evaluating group differences on the nsse and other surveys," in *annual meeting of the Florida Association of Institutional Research*, 2006, pp. 1–33.

[Sla03] M. Slater, "A note on presence terminology," *Presence connect*, vol. 3, no. 3, pp. 1–5, 2003.

[TBL+20] Z. Tang, N. J. Bryan, D. Li, T. R. Langlois, and D. Manocha, "Scene-aware audio rendering via deep acoustic analysis," *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1, 2020.

[Wil45] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945. [Online]. Available: http://www.jstor.org/stable/3001968

[WKMB16] S. Werner, F. Klein, T. Mayenfels, and K. Brandenburg, "A summary on acoustic room divergence and its effect on externalization of auditory events," in *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2016, pp. 1–6.

[ZF19] F. Zotter and M. Frank, *Ambisonics*. Springer, 2019.

[ZFSH20] F. Zotter, M. Frank, C. Schörkhuber, and R. Höldrich, "Signal-independent approach to variable-perspective (6dof) audio rendering from simultaneous surround recordings taken at multiple perspectives," *DAGA*, 2020.

[ZFZ18] M. Zaunschirm, M. Frank, and F. Zotter, "BRIR synthesis using first-order microphone arrays," in *Audio Engineering Society Convention 144*. Audio Engineering Society, 2018.

[ZSH18] M. Zaunschirm, C. Schörkhuber, and R. Höldrich, "Binaural rendering of ambisonic signals by head-related impulse response time alignment and a diffuseness constraint," *The Journal of the Acoustical Society of America*, vol. 143, no. 6, pp. 3616–3627, 2018.

[ZZF19] M. Zaunschirm, F. Zotter, and M. Frank, "Perceptual evaluation of variable-orientation binaural room impulse response rendering," in *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*. Audio Engineering Society, 2019.

# Appendices

# A Cohen's d & Cliff's Delta

In general, there are two primary strategies for the analysis of statistical data: Significance tests and effect size investigations. Significance tests suffer from two disadvantages: (I) Their significance depends on the size of the dataset, hence how many people took part in the experiment, and (II) on the cumulation of the alpha-error. This cumulation increases the chances that found significances are a coincidence. The error can be compensated by a Bonferroni-Holm correction [Hol79], and this was done in this project.

The second strategy is to investigate the effect size between two distributions. The most common effect size measure is Cohen's d [Coh13], compare equation 1. Since Cohen's d calculates a mean and a standard deviation, the investigated distributions need to follow a normal distribution. Usually, d is interpreted as follows: d = 0.2 is a small effect, d = 0.5 is a medium effect, d = 0.8 is a big effect. Still, it is important always to remember Cohen's sentence in [Coh13, p.26]:

> *We are thus reminded of the arbitrariness of this assignment of quantitative operational definitions to qualitative adjectives.*

An effect size which is applicable for all distributions is called Cliff's Delta [Cli93]. Cliff's Delta is calculated according to equation 2. It compares all samples from the first distribution with all samples from the second. All the positive differences count as +1, all the negative ones count as -1. All comparisons are summed up and normalized to obtain a delta between -1 and 1.

Note that in this thesis, we used the absolute values of both Cliff's Delta and Cohen's d.

$$\text{Cohen's d} = \frac{\bar{X}_1 - \bar{X}_2}{S}, \quad \text{with } S = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2}} \tag{1}$$

where:

| | |
|---|---|
| $\bar{x_i}$: | Sample mean of the first distribution. |
| $\bar{x_j}$: | Sample mean of the second distribution |
| $S$: | The pooled sample standard deviation. |
| $S_1^2$: | The variance of the first distribution. |
| $S_2^2$: | The variance of the second distribution. |
| $m$: | The size of the first distribution. |
| $n$: | The size of the second distribution. |

$$\text{Cliff's Delta} = \frac{\#(x_i > x_j) - \#(x_i < x_j)}{mn}, \tag{2}$$

where:

| | |
|---|---|
| $x_i$: | The ith sample from the first distribution. |
| $x_j$: | The jth sample from the second distribution. |
| $\#(\bullet)$: | Equals 1 if the argument is true. |
| $m$: | The size of the first distribution. |
| $n$: | The size of the second distribution. |

Cliff's Delta generally is a less commonly used effect size than Cohen's d. Therefore the interpretation of its results is even more difficult. It is possible to compare Cohen's d and Cliff's Delta, referred to as d and $\Delta$ from now on, for two normal distributions. Also Romano suggests this approach in [RKCS06] and obtains the following results: $\Delta$ = 0.147 is a small effect, $\Delta$ = 0.33 is a medium effect, and $\Delta$ = 0.47 is a big effect. The calculations conducted in this thesis suggest that: $\Delta$ = 0.11 is a small effect, $\Delta$ = 0.27 is a medium effect, and $\Delta$ = 0.42 is a big effect.

Figure 1 shows two normal distributions that are shifted against each other about 0, one, or four standard deviations. The results for d and $\Delta$ are plotted in figure 2 and in figure 3.

As the data in this study are composed of discrete ratings between 1 and 7, the comparison between d and $\Delta$ was also conducted for discrete data. For this purpose, two normal distributions with 38 samples were discretized to values between 1 and 7, just like the data in the listening experiment was structured. Then d and $\Delta$ were calculated for these distributions. Figure 4 shows the results: The interpretation for continuous distributions seems to be applicable also for discrete data.

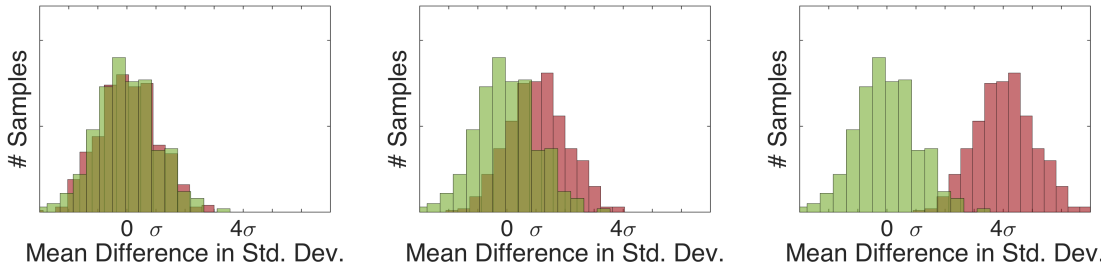



Figure 1 – Two normal distributions shifted against each other. Left: no difference between means | Middle: $\sigma$ difference between means | Right: $4\sigma$ difference between means. Both curves are composed of 500 normally distributed samples. Cohen's d and Cliff's Delta for a continuous shift between the distributions against each other are visualized in figure 2.

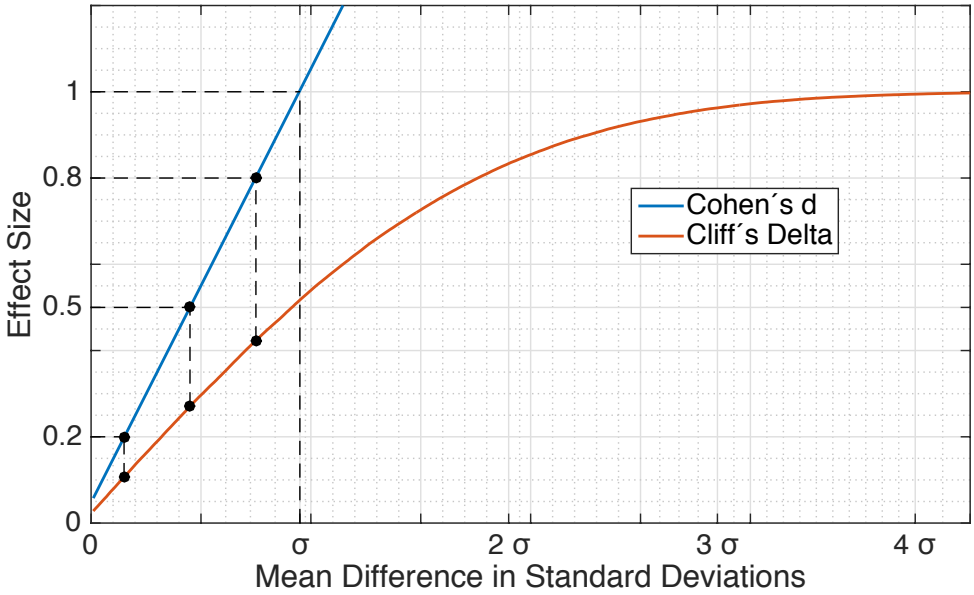



Figure 2 – Cohen's d and Cliff's Delta evaluated for two normal distributions shifted against each other. Figure 1 shows the distribution at $\sigma = 0$, 1 and 4. Figure 3 shows the same plot from 0 to $\sigma$.

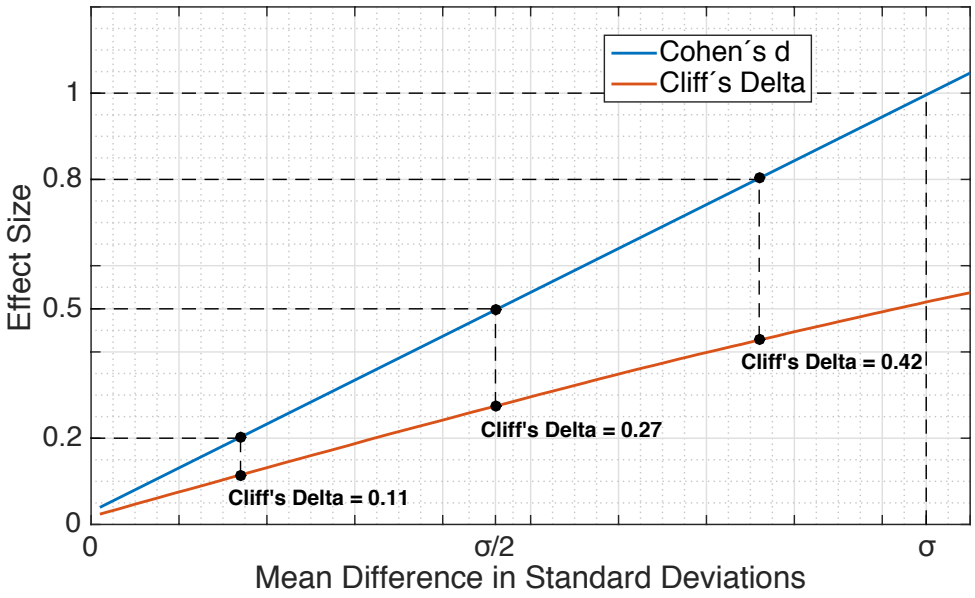



Figure 3 – Cohen's d and Cliff's Delta evaluated for two normal distributions shifted against each other. Compared to Cohen's d, Cliff's Delta values are lower for the same effect size.

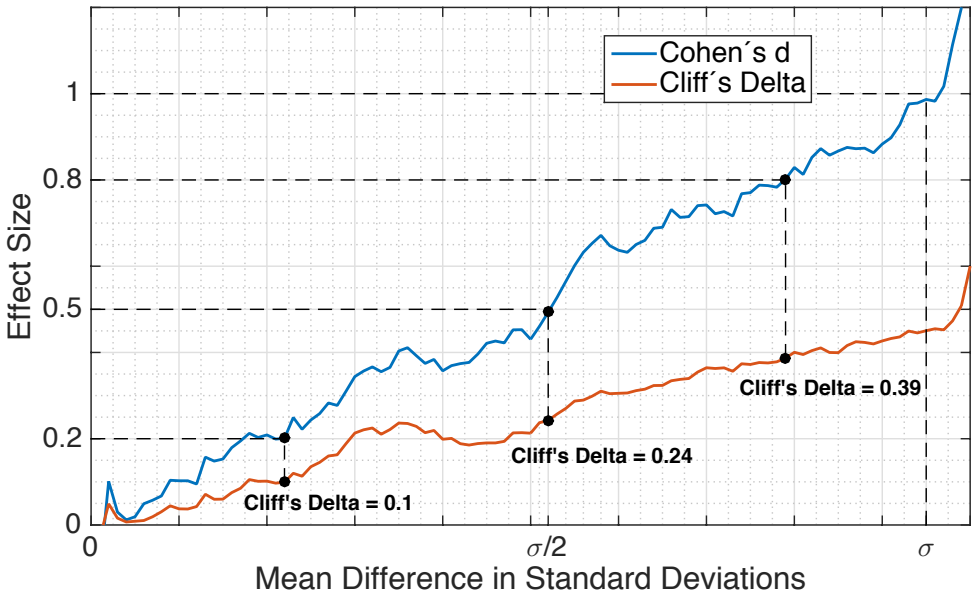



Figure 4 – Cohen's d and Cliff's Delta evaluated for two discretized normal distributions (samplesize = 38, range = 1 to 7) shifted against each other. The Cliff's Delta values from from figure 3 also model discretized distributions. A moving average filter with size 10 was used to smooth the curves in this plot.

# B C# code for Unity

This Appendix provides some central parts of the Unity C# code that was used to manage the listening experiment.

```
void Update()
  {
      OscMessage message = new OscMessage();

      message = new OscMessage();
      message.address = "/RoomEncoder/listenerX";
      message.values.Add(transform.position.z);
      osc.Send(message);
      oscPDPos.Send(message);

      message = new OscMessage();
      message.address = "/RoomEncoder/listenerY";
      message.values.Add((-1) * transform.position.x);
      osc.Send(message);
      oscPDPos.Send(message);

      if (System.Convert.ToInt32(Globals.x == 117) + System.Convert.ToInt32(
          Globals.x == 118) + System.Convert.ToInt32(Globals.x == 123) +
          System.Convert.ToInt32(Globals.x == 124) == 0)
      {

          message = new OscMessage();
          message.address = "/RoomEncoder/listenerZ";
          message.values.Add(transform.position.y);
          osc.Send(message);
          oscPDPos.Send(message);
      }

      if (System.Convert.ToInt32(Globals.x == 117) + System.Convert.ToInt32(
          Globals.x == 118) + System.Convert.ToInt32(Globals.x == 123) +
          System.Convert.ToInt32(Globals.x == 124) == 1)
      {
          message = new OscMessage();
          message.address = "/RoomEncoder/listenerZ";
          message.values.Add(transform.position.y -2.9f);
          osc.Send(message);
      }
  }
```

Listing 1 – C# Unity code for sending the position of the participant ("listener") to Reaper and to Pure Data via OSC once per frame.

```
void Update () {

          OscMessage message = new OscMessage();

        message = new OscMessage();
        message.address = "/SceneRotator/qw";
        message.values.Add(transform.rotation.w);
        osc.Send(message);
        oscPDRot.Send(message);

        message = new OscMessage();
        message.address = "/SceneRotator/qx";
        message.values.Add(transform.rotation.z);
        osc.Send(message);
        oscPDRot.Send(message);

        message = new OscMessage();
        message.address = "/SceneRotator/qy";
        message.values.Add((-1)*transform.rotation.x);
        osc.Send(message);
        oscPDRot.Send(message);

        message = new OscMessage();
        message.address = "/SceneRotator/qz";
        message.values.Add(transform.rotation.y);
        osc.Send(message);
        oscPDRot.Send(message);

    }
```

Listing 2 – C# Unity code for sending the rotation of the participant ("listener") to Reaper and to Pure Data via OSC once per frame.

```
// GET VP ANSWERS VIA THE TRIGGER
        if (triggerAction.GetStateDown(handType) && Globals.message == 0 &&
            Globals.UIActive == 0)
        {

            if (trans.rotation.eulerAngles.z < 100 && trans.rotation.
                eulerAngles.z > 60)
            {

                SendOSCMessageInt(ref oscDAW, "/track/mute", 1);
                SendOSCMessageInt(ref oscDAW, "/track/1/mute", 1);
                SendOSCMessageInt(ref oscDAW, "/stop", 1);
                WriteAnswerToJSON(Globals.i, vpWrapper, path, 1);

                Globals.i++;
                Globals.x = vpWrapper.Reihenfolge[Globals.i];
                DisplayMessage(Globals.x);
                NextStimulus(Globals.x);

                [...]

            }
        }
```

Listing 3 – C# Unity code-excerpt that manages the listening experiment. If the participant presses the trigger during a rotation of the controller between 60 and 100 degrees, then a plausibility rating of value 1 is stored and the next stimulus gets prepared.

```
void SevenOrderFullModel()
    {
        // MASTER GAIN
        SendOSCMessageFloat(ref oscDAW, "/track/8/volume/db", -7f);

        // BASS BOOST/DAMP
        SendOSCMessageFloat(ref oscDAW, "/track/8/fx/1/bypass", 0);
        SendOSCMessageFloat(ref oscDAW, "/track/8/fx/2/bypass", 1);

        // VR LS AN AUS
        SendOSCMessageInt(ref oscDAW, "/track/2/solo", 1); // 0 =

        // ECHTER LS AN AUS
        SendOSCMessageInt(ref oscDAW, "/track/1/solo", 0); // 0 =

        // Statischer Hall Kanal AN AUS
        SendOSCMessageInt(ref oscDAW, "/track/6/solo", 0); // 0 =
        SendOSCMessageInt(ref oscDAW, "/track/2/fx/2/bypass", 0);

        // RICHTCHARAKTERISTIK AN AUS
        SendOSCMessageInt(ref oscDAW, "/track/2/fx/3/bypass", 0);// mono
            convolver for eq / 0=aus, 1= an
        SendOSCMessageInt(ref oscDAW, "/track/2/fx/4/bypass", 1);// 3.Order LSP
             convolver

        // ROOM ENCODER
        SendOSCMessageInt(ref   oscRoomEncoderLS, "/RoomEncoder/numRefl", 236);
        SendOSCMessageFloat(ref oscRoomEncoderLS, "/RoomEncoder/roomX", 6);
        SendOSCMessageFloat(ref oscRoomEncoderLS, "/RoomEncoder/roomY", 4.4f);
        SendOSCMessageFloat(ref oscRoomEncoderLS, "/RoomEncoder/roomZ", 3.2f);
        SendOSCMessageFloat(ref oscRoomEncoderLS, "/RoomEncoder/reflCoeff", 0);
        SendOSCMessageFloat(ref oscRoomEncoderLS, "/RoomEncoder/
            wallAttenuationFront", -3); //-50 bis 0 db
        SendOSCMessageFloat(ref oscRoomEncoderLS, "/RoomEncoder/
            wallAttenuationBack", -9);
        SendOSCMessageFloat(ref oscRoomEncoderLS, "/RoomEncoder/
            wallAttenuationLeft", -3);
        SendOSCMessageFloat(ref oscRoomEncoderLS, "/RoomEncoder/
            wallAttenuationRight", -3);
        SendOSCMessageFloat(ref oscRoomEncoderLS, "/RoomEncoder/
            wallAttenuationCeiling", 0);
        SendOSCMessageFloat(ref oscRoomEncoderLS, "/RoomEncoder/
            wallAttenuationFloor", 0);

        // FDN REVERB
        SendOSCMessageFloat(ref oscDAW, "/track/4/volume/db", +5);
        SendOSCMessageInt(ref   oscDAW, "/track/4/mute", 0);
        SendOSCMessageFloat(ref oscDAW, "/track/4/fx/2/bypass", 1);
        SendOSCMessageFloat(ref oscFDN, "/FdnReverb/dryWet", 1f);
        SendOSCMessageInt(ref   oscFDN, "/FdnReverb/delayLength", 2);
        SendOSCMessageFloat(ref oscFDN, "/FdnReverb/revTime", 0.3f);
        SendOSCMessageFloat(ref oscFDN, "/FdnReverb/fadeInTime", 0.29f);
        SendOSCMessageFloat(ref oscFDN, "/FdnReverb/highGain", -69);

        // AMBI ORDER SETTING
        SendOSCMessageInt(ref oscDirectivityShaper, "/DirectivityShaper/
            orderSetting", 8);
        SendOSCMessageInt(ref oscRoomEncoderLS, "/RoomEncoder/orderSetting", 8)
            ;
        SendOSCMessageInt(ref oscSceneRotator, "/SceneRotator/orderSetting", 8)
            ;
        SendOSCMessageInt(ref oscBinauralDecoder, "/BinauralDecoder/
            inputOrderSetting", 8);
        SendOSCMessageInt(ref oscSceneRotator90, "/SceneRotator/orderSetting",
            8);
    }
```

Listing 4 – C# Unity code to prepare Reaper for the next stimulus whenever it will be the 7th-order Ambisonics model.

# C IEM Plug-in GUIs

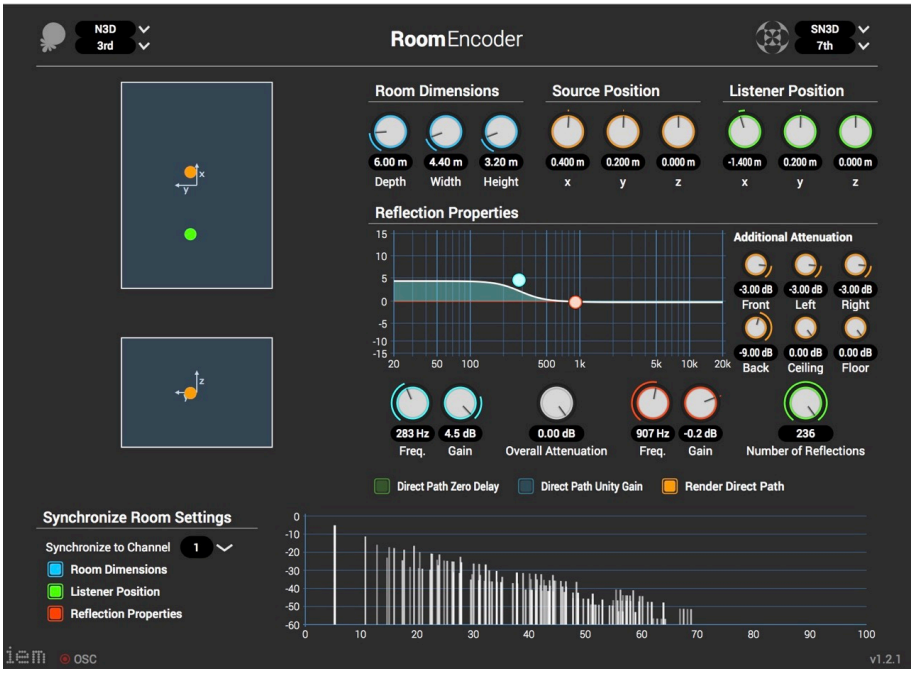



Figure 5 – The Room Encoder Plug-in with the parameter settings for the 7th order model.

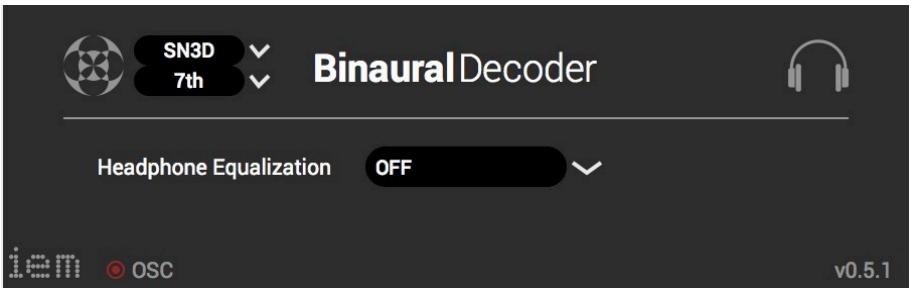



Figure 6 – The Binaural Decoder Plug-in with the parameter settings for the 7th-order model.

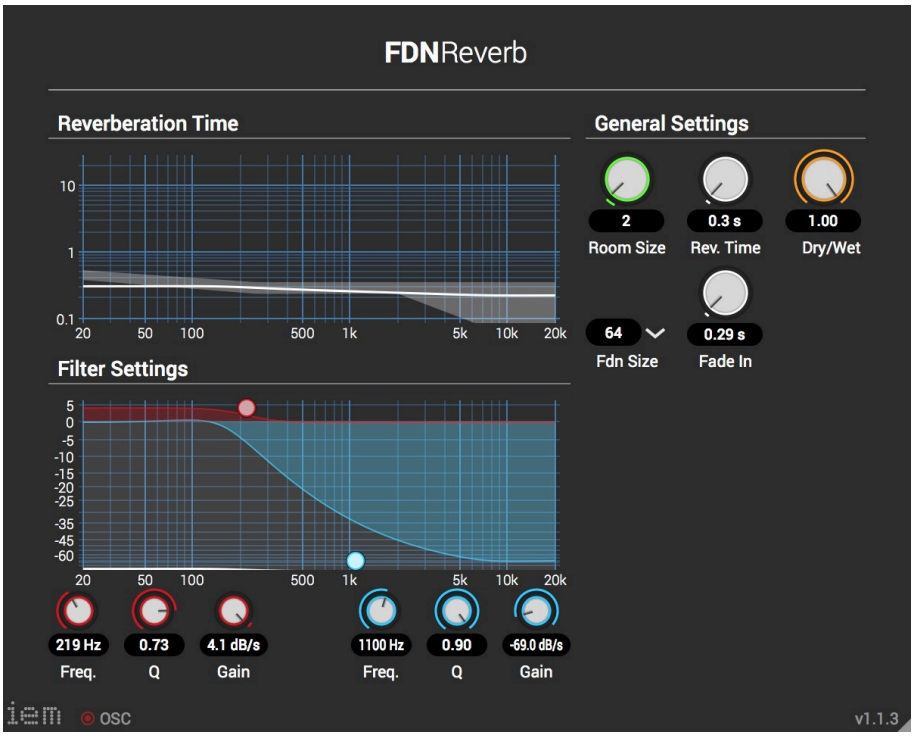



Figure 7 – The FDN Plug-in with the parameter settings for the 7th order model.

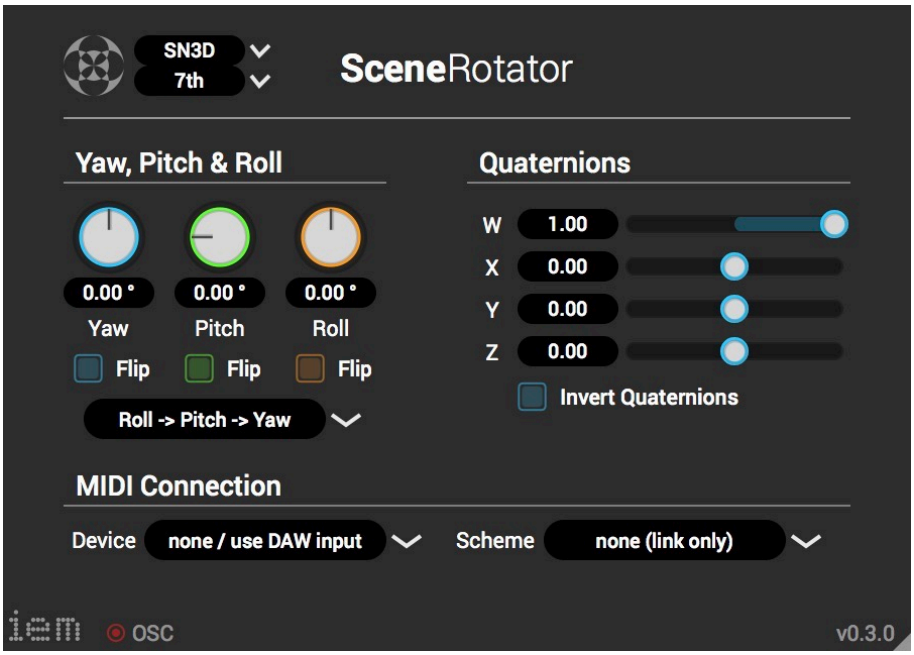



Figure 8 – The Scene Rotator Plug-in for the 7th-order model.

# D Original experiment-instructions

The three next pages provide the original experiment-instructions given to the participants in german language.

# 🎮 🎧 VR_IEM: EIN HÖRVERSUCH in VIRTUAL REALITY 🎮 🎧

Willkommen im virtuellen IEM und herzlichen Dank für die Teilnahme an diesem Hörversuch. Bitte lesen Sie diese Erklärung in Ruhe durch. Dieser Versuch wird etwa eine Stunde dauern, und wir werden zwei kurze Pausen machen.

In dem nun folgenden Hörversuch werden Sie gebeten, sich mittels Virtual Reality Brille und Kopfhörern in verschiedene virtuelle Räume zu begeben. In diesen Räumen werden Sie einen virtuellen Lautsprecher sehen und hören können. Der Klang wird dabei binaural über Ihre Kopfhörer wiedergegeben. Ihre Aufgabe in den virtuellen Räumen wird es sein die Plausibilität der Darstellungen zu bewerten.



Bei der Bewertung der Plausibilität sollen Sie sich im Allgemeinen folgende Fragen stellen:

1. „Ist DIESE audiovisuelle Darstellung plausibel?“
2. „Passt das was ich HÖRE zu dem was ich SEHE?“
3. „Könnte ein Lautsprecher in einem SOLCHEN Raum SO klingen?“

Bitte achten Sie bei Ihrer Entscheidung besonders auf folgende Parameter:

1. Externalisierung: Fragen Sie sich: „Klingt der Klang so, als wäre der virtuelle Lautsprecher außerhalb meines Kopfes?“
2. Lokalisation: „Höre ich den virtuellen Lautsprecher aus der Richtung, in der ich ihn sehe?“
3. Distanz: „Stimmen die akustische und die visuelle Distanz zum Lautsprecher überein, oder klingt er näher/ferner als er aussieht?“
4. Klangfarbe: „Ändert sich die Klangfarbe des Lautsprechers abhängig von meiner Position im Raum plausibel/meinen Erwartungen entsprechend?“ - Bedenken Sie dabei, dass ein realer Lautsprecher nicht aus allen Richtungen gleich klingt.
5. Nachhall bzw. Raumeindruck: „Stimmen mein akustischer Raumeindruck und mein visueller Raumeindruck überein? Klingt der Raum so groß wie er aussieht?“

Für die absolute Bewertung der Plausibilität wird Ihnen eine siebenteilige Skala zur Verfügung stehen:

- 7… Völlig plausibel - bedeutet: „Akustische Authentizität ist möglich = Es wäre möglich, dass ein ECHTER Lautsprecher in einem SOLCHEN Raum SO klingen würde.“
- 6
- 5
- 4 … Weder noch - bedeutet: „Ich erkenne, dass die Virtualisierung nicht hervorragend passt, finde sie deswegen aber nicht völlig unplausibel“
- 3
- 2
- 1 … Völlig unplausibel - bedeutet: „Der Klang und das Bild passen überhaupt nicht zusammen, oder zumindest einer der oben genannten Faktoren weicht völlig von meinen Erwartungen ab.“

Wenn Sie bei einem Stimulus eine Entscheidung getroffen haben, wird Ihre Antwort gespeichert und die Wiedergabe pausiert. Sie sehen dann die Instruktionen bezüglich Ihrer Bewegungsfreiheit für das nächste Beispiel.

Diese Instruktionen werden Ihnen nach jedem Stimulus angezeigt und können wie folgt lauten:

- Gehen Sie zu Punkt A | Kopf NICHT bewegen
- Gehen Sie zu Punkt A | Kopf BEWEGEN
- Gehen Sie frei im Raum | NICHT interagieren
- Gehen Sie frei im Raum | INTERAGIEREN

Wenn Sie die jeweilige Instruktion gelesen haben, gehen Sie gegebenenfalls zum Punkt A zurück und drücken Sie danach auf Play, um den nächsten Stimulus zu starten. Als Interface benutzen Sie einen VR-Controller.

Vor dem Hörversuch werden Sie eine kurze Trainingsphase durchlaufen, in der Sie die Möglichkeit haben sich an die virtuelle Umgebung zu gewöhnen. In der Trainingsphase werden Ihnen schon einige der später vorkommenden Stimuli präsentiert.
Ziel der Trainingsphase ist es ein erstes inneres Raster der Plausibilitäten aufzubauen, das Sie später für Ihre Entscheidungen nutzen können. Während des Versuches werden Sie sich auf der absoluten Skala (1-7) für Ihre Antwort entscheiden, ohne zwischen den Beispielen hin und her zu wechseln. Ihr inneres Raster wird sich im Laufe des Versuchs natürlich weiterentwickeln, die Trainingsphase soll Ihnen den Einstieg in den Hörversuch erleichtern.

Ein Stimulus setzt sich immer aus drei Kategorien zusammen, nämlich aus:

A. Den Videodarstellungen: Zu sehen wird sein…
   1. Nichts = schwarzes Bild in der VR Brille, oder
   2. Einer von vier virtuellen Räumen, oder
   3. Die Realität ohne VR Brille.

B. Den Audiodarstellungen entsprechen akustischen Virtualisierungen des Lautsprechers in verschiedener Qualität. So soll untersucht werden, welche Parameter der akustischen Virtualisierung für eine plausible Darstellung besonders wichtig sind.

C. Ihrer Bewegungsfreiheit:
   1. Statisch: Sie stehen an einer fixen Position A und sollen den Kopf während des Hörens nicht bewegen. Schauen Sie dabei in Richtung des Lautsprechers.
   2. Kopfrotation: Sie stehen wieder an einer fixen Position A, dürfen den Kopf aber drehen.
   3. Gehen: Sie dürfen sich frei im Raum bewegen, sollen aber keinen physischen Kontakt zu den Wänden oder der Einrichtung aufnehmen.
   4. Interagieren: Sie dürfen sich frei im Raum bewegen und sollen mit dem Raum (Wände, Tische) in physikalischen Kontakt treten - beispielsweise in dem Sie sich an einer Wand anlehnen oder sich auf einen Tisch setzen.

Nach einem Drittel und nach zwei Drittel des Versuchs werden wir kurze Pausen machen, um die Ohren und Augen zu entspannen.
Alles klar soweit?!?! In der Trainingsphase werden sich die meisten Fragen beantworten! Wechseln wir nun ins virtuelle IEM, los gehts!!