DEGREE PROJECT

Comparison between binaural reproduction methods for auditory distance perception in virtual reality

Filip Brihage

Audio Technology, bachelor's level
2019

Luleå University of Technology
Department of Arts, Communication and Education

LULEÅ UNIVERSITY OF TECHNOLOGY

Comparison between binaural reproduction methods for auditory distance perception in virtual reality

by

Filip Brihage

Bachelor's Thesis

Audio Technology
2019
Department of Arts, Communication and Education
School of Music in Piteå

Although a large amount of work is available, our knowledge of distance hearing is deficient compared to our knowledge of directional hearing. This deficiency is due to the extraordinary complexity of the subject.

(Blauert, 1997, p.117)

Abstract

In recent years, virtual reality (VR) has become more common within the media industry and with its increasing popularity, spatial audio reproduction methods such as ambisonics have got more attention. But how humans perceive auditory events in VR is an ongoing research topic and there is a need for keep evaluating the reproduction methods used for it.

Even though humans' auditory distance perception in natural and laboratory conditions has been investigated in many previous studies, little has been made in VR environments. Thus, there is a need for keep investigating how humans estimate sound source distances in VR and which recording techniques that results in most accurate estimations.

This study aimed to investigate human's accuracy in sound source distance estimations in VR, when sound sources have been recorded with an artificial head or first-order ambisonic (FOA) microphone. Three different reproductions methods were used in the study: artificial head, FOA-tracked binaural and FOA-static binaural. In a VR environment, twenty-three subjects were asked to position a virtual loudspeaker at the same position as a sound source. The results showed no significant differences between the three reproduction methods. In the results it was also shown that the subjects clearly overestimated the sound source distances. A possible reason is that the subjects underestimated the size of the virtual environment.

Preface

During the work resulting in this thesis, I have been lucky to get help from many people. I would like to thank my classmates and friends Anton Louthander, Jesper Thiman and Daniel Fröjd Wasberg for their participation in the pre-study, helping me improve the experiment. Thanks to Tyréns AB for loan of an HTC VIVE headset and ambisonic microphone. Thanks goes also to Nyssim Lefford and Jan Berg for helping me shape my thinking about how to conduct and present an experiment like this. I would also like to thank my long-time friend Theodor for helping me design the VR environment. A special thanks is given to my classmate and friend Markus Bélteky for all his help during my work. Thanks for performing the monologue, participating in the pre-study, assisting at the recording of the stimuli and lending me an Xbox controller. Finally, I would like to thank my supervisor at LTU, Arne Nykänen, for all his support and hours of conversations guiding me through difficult parts of the work.

Filip Brihage
Piteå, Sweden
April 9, 2019

Contents

List of abbreviations

6-DOF	Six degrees of freedom
ANOVA	Analysis of variance
FOA	First-order ambisonic
FOV	Field of view
HMD	Head-mounted display
HRTF	Head-related transfer function
IHL	Inside-the-head locatedness
ILD	Interaural level difference
IPD	Interpupillary distance
ITD	Interaural time difference
SPL	Sound pressure level
VR	Virtual reality

CHAPTER 1

Introduction

By using head-mounted displays (HMD) media technology developers can create illusions of visually being enveloped by virtual worlds and with new gear, virtual reality (VR) has in recent years become more popular within the media industry. But a modern VR system does not only consider the visual sense in order to make the user believe she is placed in a virtual world, it also considers the haptic and aural senses. Even though there is no actual virtual world outside the user's field of view (FOV), sounds can help developers make the user believe it. Thus, audio can constitute a key feature to make the VR user believe she is part of a virtual world.

When the VR user turns her head, 3D audio (full-sphere surround) needs to be considered for the sounds to continuously follow along with the virtual visual content. But how to design audio for VR is an ongoing research topic and within the scope of 3D audio there are two primary factors to consider. The first is how we perceive the direction of a sound source (i.e. the VR user's ability to tell where a source of an auditory event is coming from). The second is how we perceive sound source distance (e.g. the VR user's ability to estimate whether a sound source is located close or far away). How we aurally perceive distance in VR is foundation for the question investigated in this study.

1.1 Head tracking

In order to make a VR user believe she is placed inside a virtual world, it is important that the visual content continuously changes based on the movement of the user's head (i.e. the movement of the VR headset). This is a key feature for VR applications called 'head-based displays' (Sherman, & Craig, 2003) and is one important element that differentiates it from other mediums. In six degrees of freedom (6-DOF) situations both the position

and orientation of the user's head is reported to the computer using head tracking. And when an auditory event also changes with the user's movements, we can define VR as "… a high-end user-computer interface that involves real-time simulation and interactions through multiple sensorial channels." (Burdea, & Coiffet, 2003, p.3).

Regarding auditory distance perception in VR environments, it has not been thoroughly investigated whether head movements improve our abilities to estimate sound source distances or not. But in a real-world experiment by Simpson and Stanton (1973) head movements was shown to not improve distance estimations. There is however space for further investigating this in VR environments.

1.2 Binaural recording technique

One way to simulate how we perceive auditory events in the real world is to make binaural recordings using an artificial head. The typical artificial head consists of two omnidirectional capsules positioned in artificial ears which have been mounted on an enclosure, built like the shape of a human head. The main idea with the technique is during playback of an auditory event reproduce the spatial cues the listener would have been provided in the source environment (Rumsey, 2017a). Thus, by listening to a binaural recording through headphones the user is supposed to perceive the auditory event just as she would have had in the source environment.

One issue with such reproductions is that every individual has a unique head-related transfer function (HRTF). Due to the position and shape of the pinnae, head, shoulders, etcetera, every individual receive sound waves a little bit different. And as we grow up our brains learn how to estimate the location of a sound source due to how the sound waves arrive to our ears and thus, it might be hard for a VR user to estimate the location of a sound source using a generic HRTF. But how important our individualized HRTF is in a VR application with head tracking is still vague (Rumsey, 2017b).

When producing audio for VR it is beneficial if recorded sounds can be converted into 3D audio in order to be used with head tracking. Unfortunately, there is no straightforward method to reproduce artificial head recordings with head tracking (Hong, Lam, Ong, Ooi, Gan, Kang, Feng, & Tan, 2019). That is, an artificial head recording does not contain enough spatial information to calculate new HRTF signals. But since an artificial head is designed for precisely simulate how we perceive auditory events it would however be interesting to, in a static mode, compare it to other recording techniques normally used for VR applications.

1.3 First-order ambisonic (FOA) recording technique

First-order ambisonics (FOA) is a technique for recording and reproducing sounds in 3D and can be described as a 3D version of the MS stereo technique with added channels for depth and height. The typical FOA microphone consists of four sub-cardioid (between cardioid and omni) capsules positioned in a tetrahedral arrangement (Rumsey, 2017a). These four capsules point in different directions and correspond to left-front (points upwards), right-front (downwards), left-back (downwards) and right-back (upwards). Thus, the A-format of the FOA system involves four recorded signals which can, using sum and difference techniques, be decoded to a B-format signal.

"The B-format consists of four signals that between them represent the pressure and velocity components of the sound field in any direction…" (Rumsey, 2017a, p.113) and they are called W, X, Y and Z. The W channel is obtained by in phase adding all four outputs of the capsules together and represents the omnidirectional pressure component. The X, Y and Z channels are instead derived by utilizing three different difference techniques and they all represent figure-eight components facing different directions (forward, sideways and upwards) in the sound field.

The B-format of ambisonics allows for a dynamic binaural playback through headphones by applying HRTFs to the signal. With the use of head tracking the recorded sound field can then be rotated with the movement of the HMD (Thresh, Armstrong, & Kearney, 2017). Since B-format ambisonics can reproduce sound fields in 3D, the technique is popular for designing audio for VR applications and 360-degrees videos.

1.4 Inside-the-head locatedness (IHL)

When reproducing recorded audio over headphones, there is risk for inside-the-head locatedness (IHL). That is, even though a sound source has been recorded 10 m away from the microphone, the listener perceives the reproduced audio as located inside the head. Why this effect occurs has been investigated in many previous studies and inaccurately reproduced HRTF signals have been pointed out as one possible reason (Hur, Park, Lee, & Young, 2008). According to Blauert (1997) IHL may also occur when audio is reproduced over headphones and both ears receive identical or highly similar signals.

In Hur et al. individualized HRTFs was shown to help listeners perceive recorded audio as more externalized rather than inside their heads. However, if one of the methods compared in the present study generate a generic HRTF which translates well to the different individualized HRTFs of the participants, then that method might benefit from it, produce less IHL

effects, and result in more accurate sound source distance estimations in the VR environment.

While in theory, individualized HRTFs might be the optimal solution for adequate externalization of auditory events when reproduced over headphones, measuring each VR user's individualized HRTF is in practice cumbersome. It requires specialized equipment and takes a lot of time. Creating an optimal generic HRTF is therefore an ongoing research topic.

1.5 Thesis aim and research questions

The present study seeks to answer if there is, in a VR environment, any difference in how accurately we aurally estimate sound source distances when the source has been recorded with an artificial head or FOA microphone. And if there is a difference, which technique result in most precise estimations? To answer these questions the main research question for the present study is: what is the auditory precision in perceived distance to binaurally reproduced sound sources at different distances in a VR environment, when recorded with an artificial head or a FOA microphone?

Even though an artificial head recording, as previously mentioned, cannot be used for 3D audio, it is still interesting to investigate whether the technique result in more accurate auditory distance estimations than FOA. Does FOA already outperform the artificial head in making us accurately estimate sound source distances in VR, or does the FOA technique still need improvements? The FOA technique is after all, in contrast to the artificial head, a computer-generated way of reproducing how we perceive auditory events.

Another topic of the present study is to investigate whether head movements improve our abilities to estimate sound source distances in VR. Therefore, the present study will try to compare FOA-tracked binaural (i.e. headphone-based head tracked binaural) with FOA-static binaural reproduced audio regarding how accurately we estimate sound source distances in VR.

Finally, since VR is a relatively new platform for most sound designers there is a need for evaluating the recording techniques for capturing sounds for VR. Thus, the general aim of the present study is to contribute towards deep knowledge about how to record and design audio for VR.

CHAPTER 2

Background

Even though distance perception in VR environments has not been thoroughly investigated, there are many previous studies regarding auditory distance perception, with and without visual cues, conducted in natural and laboratory conditions. This section will try to highlight some of the most important previously investigated factors to consider, when conducting an VR experiment regarding auditory distance perception.

2.1 Distance perception: auditory cues

This section provides information about important factors regarding how we perceive auditory distance.

2.1.1 Intensity

To estimate the distance of a sound source the perceived intensity (interpreted as loudness) of the source is an important auditory cue (Begault, 1994). And in theory, the sound intensity of a point-source in a free field is attenuated by 6 dB for each doubling distance (Everest, & Pohlmann, 2015). Thus, we should theoretically perceive a sound source 2 m away as 6 dB quieter than the same source at 1 m. Though this is only true in a free field, it has been shown that even in reverberant conditions the sound intensity of sound sources is attenuated for each doubling distance (Begault). By experiencing this interrelation between increasing distance and intensity attenuation in many visual-aural events, we improve our abilities to estimate sound source distances. This fact could indicate that users of VR will be better at estimating virtual sound source distances the more they have experienced VR applications.

2.1.2 Reverberation

When recording stimuli for an experiment studying auditory distance perception it is important to consider the acoustic properties of the recording environment. Recording stimuli in an anechoic chamber is for instance not a good idea since we have very hard to estimate auditory distances without reverberation cues. A study by Mershon & King (1975) showed that the "…intensity of a sound can serve only as a relative cue to changes in egocentric auditory distance." (p.413). That is, we need reverberation cues to be able to accurately estimate the distance of a sound source. The present study will therefore record the auditory stimuli in a reverberant space.

2.1.3 Familiarity

Since we improve our abilities to estimate sound source distances by experiencing many different visual-aural events, our extent of familiarity with a certain auditory event should affect our ability to estimate the sound source distance in that context. In an experiment, Gardner (1969) showed that even though the subjects were equally familiar with shouts and whispers, they clearly overestimated the distance of shouts, while underestimating the distance of whispers. Based on the results of the study, a virtual buzzing mosquito rendered 5 cm away from the VR user should be easier to correctly interpret than a mosquito rendered at 5 m. This is due to that we are more familiar with hearing a mosquito 5 cm away. Because even though VR developers might have the tools for positioning a buzzing mosquito far away from the user, they must consider the fact that most people have not experienced such an auditory event before. Therefore, based on previous experience hearing mosquitos, the VR user might perceive the mosquito much closer than the developer intended. The present study will therefore, as stimuli, present a virtual loudspeaker together with recordings of spoken sentences. The main reason is that the subjects are expected to have experienced such an auditory event at many different distances. Example of such situations could be watching tv or hearing a PA announcement.

2.1.4 Binaural cues

To describe the location of a sound source relative to a listener, previous literature has used a head-related system of coordinates (Blauert, 1997). One coordinate axis is called the median plane and vertically divides the human body in two exact halves. "If the head may be assumed symmetrical, it is then symmetrical about the median plane." (Blauert, p.14). Thus, when a sound source is in the median plane, both ears theoretically receive identical input signals. But when conducting an experiment involving head tracking,

non-identical ear input signals will also be evaluated. When the VR user turns her head, the ears will no longer receive identical input signals and the auditory event will not occur in the median plane.

When a sound source is located somewhere to the left or right of the median plane, two binaural cues are important in order to estimate the direction of the source: the interaural time differences (ITDs) and the interaural level differences (ILDs). ITD is when our brain utilizes the difference in time a signal arrives to our ears to estimate the sound source direction, while ILD is when our brain uses the difference in sound pressure level (SPL) a signal inputs to our ears in order to estimate the sound source direction.

In contrast to a FOA-static binaural or artificial head recording when the sound source has been recorded in the median plane, FOA recordings with head movements (i.e. FOA-tracked) can result in use of binaural cues. That is, both ears may not receive the same input signals anymore. Thus, conducting an experiment evaluating possible differences between FOA-tracked and FOA-static binaural will inevitably try to evaluate how important ITD and ILD are for auditory distance perception.

2.2 Visual-aural relationship

This section provides information about important factors regarding how we can be affected by visual cues when estimating sound source distances.

2.2.1 Ventriloquism effect

When we experience a visual-aural event, the 'ventriloquism effect' can make us misjudge the location of the sound source (Vroomen, & De Gelder, 2004). The effect can occur when visual and aural cues are presented to us in close temporal and spatial context. In such situations our perceptual system interprets the two different cues as one single event has occurred and attempts to decrease the conflict between the locations of the visual and aural cues. And since our spatial resolution is generally more accurate in the visual modality than the auditory (Vroomen, et al.), it seems more logical for our brain to adjust the location of the aural cue. One example of a visual-aural event when this effect typical occurs, is during the performance of a ventriloquist. Even though the sound source location is the body of the ventriloquist, movements of the puppet's mouth makes the audience perceive the source location as the puppet.

In Boland et al. (2012) the ventriloquism effect was shown to affect the subject's estimations of sound source distances. The results indicate that one can create a considerable mismatch between the position of a visual and aural cue, we will still perceive the visual-aural scene as consistent. In an attempt to avoid the ventriloquism effect in the present study, the test

subjects will be asked to position a virtual loudspeaker at the position they perceive as the location of the sound source.

2.2.2 Proximity-image effect

One effect related to the ventriloquism effect is the 'proximity-image effect'. It was in an experiment by Gardner (1968) shown to affect subject's perception of the location of a sound source. In an anechoic room, Gardner positioned several loudspeakers in front of the listening position and since the loudspeakers were positioned directly behind each other, only the closest one was visible for the subject. It was then showed during listening tests, that no matter which loudspeaker played back a recorded speech, the closest loudspeaker was always chosen as the sound source. Even though this experiment was conducted in an anechoic room, it may indicate that subjects in a study regarding auditory distance perception in VR environments, will be particularly affected by a close-up visual cue.

CHAPTER 3

Method

The present study investigates our accuracy in sound source distance estimations in VR, when the sound sources have been recorded with an artificial head or FOA microphone. In an experiment, subjects were asked to position a virtual loudspeaker at the same position as invisible sound sources. This chapter explains how the stimuli were produced, the test environment was designed and the listening tests were conducted.

3.1 Auditory stimuli

3.1.1 Recording

To design the test environment, auditory stimuli were pre-recorded in a reverberant hall (the dimensions of the hall were approximately the same as the virtual hall, see fig. 1). During the recording session, an auditory source was reproduced by a Genelec 1030A loudspeaker at five distances from the microphone position (2 m, 3 m, 5 m, 7 m, 10 m). Positioned on a speaker stand, the middle of the bass driver was situated 145 cm above the floor and the SPL was approximately 65 dBA 1 m from the loudspeaker. The auditory source was a short monologue in Swedish recorded by a male speaker. One at a time, two different microphones (Neumann KU 100 & Soundfield SPS 200) were positioned at the microphone position at the same height as the loudspeaker and an RME Octamic II was used to amplify the microphone signals. To avoid variances in amplification level for the different microphone capsules, all gain levels were set to a maximum of 60 dB. An RME Babyface was then used to record the signals into Pro Tools in 48 kHz, 24 bits. Since five recordings were made for each microphone technique, a total of ten auditory stimuli were recorded.

3.1.2 Reproduction Methods

The five FOA recordings were first converted into B-format FuMa using the plug-in 'Surround Zone 2' (Soundfield, 2019) and then into B-format ambiX using the plug-in 'Ambi Converter' (Noisemakers, 2019). In the game engine Unreal Engine 4 (Epic Games, 2019), using the spatial audio plug-in 'Resonance Audio' (Resonance Audio by Google, 2019), the B-format ambiX files were then binaurally decoded for headphone playback with head tracking. The artificial head recordings were played back unprocessed in the game engine. All ten recordings were converted into 44,1 kHz, 16 bits by recommendation of Epic Games (2019).

The FOA recordings were copied and used for both FOA-static binaural and FOA-tracked binaural. When the FOA-tracked binaural stimuli were played back in the game engine, an informative widget encouraged the participants to move their heads. When the FOA-static binaural and artificial head stimuli were played back in the game engine, a widget informed the participants to keep their heads still.

3.1.3 Calibration

During the recording session, the artificial head microphone amplified the input signals more than the FOA microphone. When imported into the game engine, the integrated LUFS for each auditory stimuli pair (for 2 m, 3 m, etcetera.) was therefore matched by adjusting the 'volume multiplier' function. The integrated LUFS was measured by recording the output of the game engine into Pro Tools using the audio application 'Voicemeeter' (VB-Audio Software, 2019). The different audio files were in Pro Tools measured by using the plug-in 'Insight' (Izotope, 2019).

Table 1: Final measured LUFS-values and volume multiplier settings in the game engine

Audio Stimulus	Volume Multiplier	Integrated LUFS
Ambisonics 2 m	1	-32.7
Artificial head 2 m	0.6	-32.7
Ambisonics 3 m	1	-35.8
Artificial head 3 m	0.595	-35.8
Ambisonics 5 m	1	-38.5
Artificial head 5 m	0.6	-38.5
Ambisonics 7 m	1	-40.1
Artificial head 7 m	0.595	-40.1
Ambisonics 10 m	1	-41
Artificial head 10 m	0.61	-41

3.2 Test environment

3.2.1 Pre-study

In order to get feedback regarding the listening test design and setting an adequate listening level, four audio technology students participated in a pre-study. Each student performed a complete version of the listening test by reading the written instructions and performing the tasks in the VR environment. Both during and after the listening tests, they were individually asked about their thoughts regarding the test procedure. When the pre-study was completed, the written instructions and virtual setting had been updated according to the students wishes. An adequate listening level had also been set.

3.2.2 Virtual setting

To avoid an obvious discrepancy between the auditory and visual stimuli in the virtual environment, the approximate size and shape of the recording hall was virtually designed in the game engine. For mainly two reasons visual cues were implemented into the virtual hall. In Zahorik (2001) it was shown that we with help of "visual anchors" have easier to estimate the distance to sound sources. And by placing visual cues in the virtual hall, the ecological validity of the present study increases in the sense that VR applications rarely consists of empty environments.

In the game engine, the virtual loudspeaker and camera position (i.e. the listening position) were rendered at the same height as in the recording hall (during the listening tests, the height of the camera position varied a bit depending on the height of the participant). The auditory stimuli were positioned at the same height and distances from the camera position as they were recorded from the microphone position in the recording hall (i.e. 2 m, 3 m, 5 m, 7 m, 10 m). To accomplish this, the system of measurement in the game engine was used were 1 'Unreal Units' represents 1 cm (Epic Games, 2019).

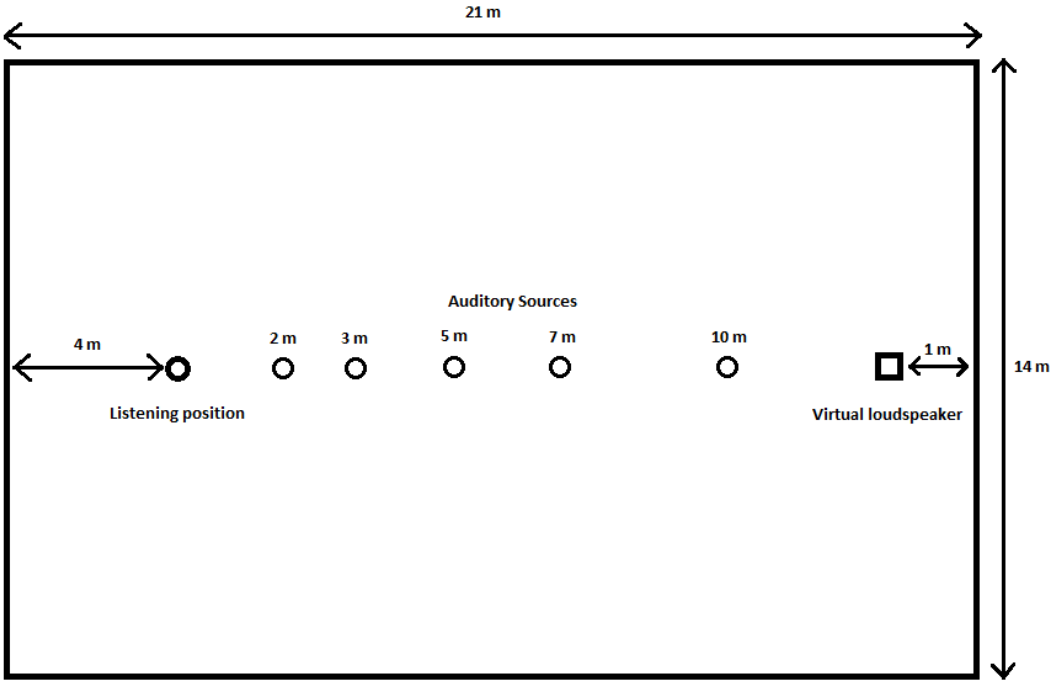
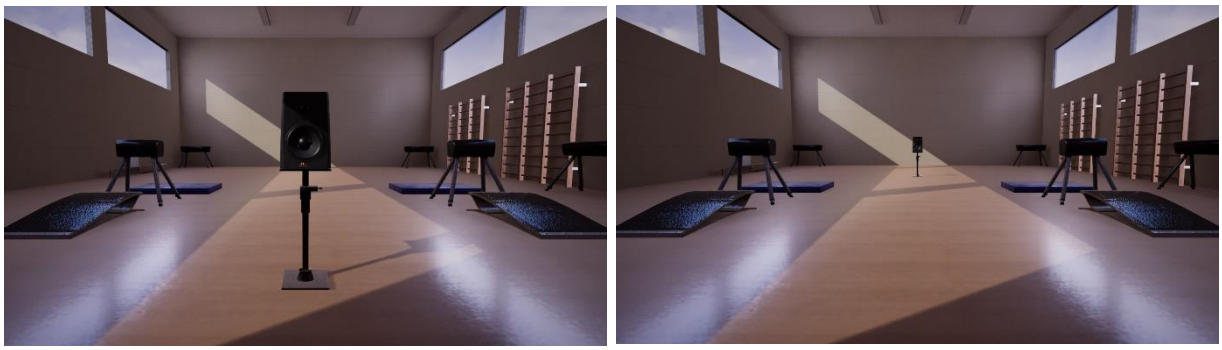


Figure 1: Dimensions of the virtual hall, (height: 7m).



(a) (b)

Figure 2: Virtual hall created for the tests. Loudspeaker rendered at 2 m (a) and 10 m (b).

Figure 3: Positions of the auditory stimuli (grey speaker symbols).

As the visual target, a loudspeaker was chosen in order to take advantage of the participants' trust in loudspeakers as devices which generate sounds. Furthermore, in previous studies regarding auditory distance perception loudspeakers have been used as visual targets (Paquier, Côte, Devillers, & Koehl, 2016).

3.2.3 Real-world setting

The listening tests were carried out in a class room at the School of Music in Piteå. The participants were positioned on a chair facing away from the experimenter and the auditory stimuli were played back over closed-back studio headphones from AKG (K 272 HD). The tests were conducted using an HTC VIVE headset and tracking system and an Xbox controller was used to navigate in the virtual environment.

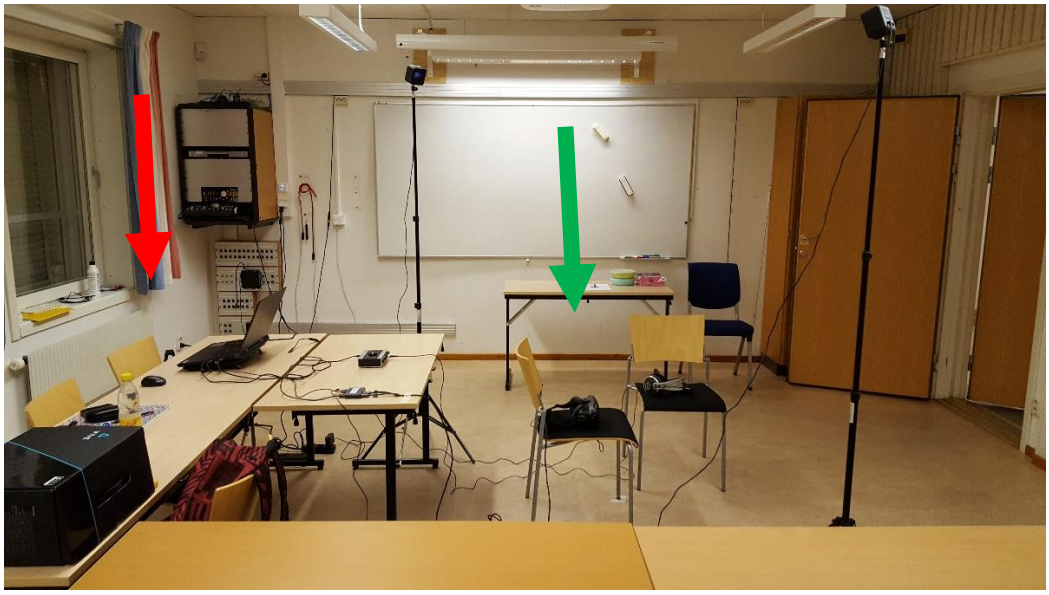


Figure 4: School of Music in Piteå: (red arrow) position of the experimenter, (green arrow) position of the subject.

3.3 Listening test

To measure the participants auditory precision in perceived distance to sound sources in VR, they were asked to position a virtual loudspeaker at the same position as an invisible sound source. The auditory stimulus was in real time, randomly chosen from the fifteen stimuli and played back by the game engine (no method was used to guarantee that the same order did not occur for more than one participant). The participants were limited to hear each stimulus one time. This limitation was utilized to force the participants to provide spontaneous distance estimations. During the listening tests, a total of fifteen trials were conducted per subject and since the present study

employs a within-subjects design, each subject was presented with the same fifteen stimuli.

To position the loudspeaker, the participants were able to stepless move it along the x-axis (closer or further away from the camera position) by using two buttons on the controller. The participants also used the controller to confirm the new position of the loudspeaker, stop the audio (if still playing) and move on to the next trial. For each trial, the loudspeaker spawned at the same position (see fig.1) and the participants were not forced to listen to the entire auditory stimulus. The fifteen scores for each participant were automatically saved by the game engine.

3.4 Procedure

The subject was first provided with a written description of the experimental task. After reading the instructions, the subject was asked to sit down and put on the headset. Due to convenience, the experimenter then helped the subject with putting on the headphones and locating the controller.

During the test, the subject was not allowed to alter the listening level. The subject was however, to have an optimal sight, allowed to adjust the interpupillary distance (IPD) by using the 'IPD knob' on the headset.

Before the actual test, there was a familiarization part in the VR environment where the subject answered a question about previous experience of VR applications. Then the subject was able to look around in the hall for 20 seconds and was then, in order to move on to training trials, supposed to move the loudspeaker to a mark on the floor. During the familiarization part no audio was played, and the subject could ask the experimenter about the game mechanics.

After the familiarization part, the subject performed two training trials. Both auditory stimuli in this part were randomly chosen from the fifteen stimuli for the actual test. The scores from the training trials were not saved. After the two training trials, the actual listening test begun.

A total of twenty-three subjects participated in the listening tests. They were all media or music students at the School of Music in Piteå and reported no auditory impairments. All tests were conducted in Swedish.

CHAPTER 4

Results

The relative error E between the perceived distance (d_{per}) and the actual distance of the sound source (d_{src}) was for each score, calculated using the following equation:

$$E = \left|\frac{(d_{per} - d_{src})}{d_{src}}\right| \quad (1)$$

Then every subject's mean relative error for each of the three experimental conditions was calculated using the following equation:

$$\bar{E} = \frac{E_{2m} + E_{3m} + E_{5m} + E_{7m} + E_{10m}}{5} \quad (2)$$

A repeated-measures analysis of variance (ANOVA) was then carried out.

Since the Mauchly test showed that the assumption of sphericity was not violated, $\chi^2(2) = 2.48$, $p > .05$, no correction method was applied to the result. With a one-way repeated-measures ANOVA no significant differences were found between the experimental conditions, $F(2, 42) = 0.66$, $p > .05$.

Out of 345 scores ($15 * 23$), 340 was used in the analysis. For one subject, all artificial head scores were excluded since the subject perceived all those sound sources as located behind the head. During the listening test the subject and experimenter agreed on that the subject, when perceiving an auditory event from behind, would position the loudspeaker as close to the listening position as possible. Afterwards, it was in the scores shown to be all five artificial head stimuli.

Table 2: Mean relative error and standard deviation for each experimental condition

Experimental condition	Mean relative error	Standard Deviation
FOA-tracked binaural	0.72 (72%)	0.44
Artificial head	0.68 (68%)	0.42
FOA-static binaural	0.79 (79%)	0.46

As can be seen in table 2, the experimental conditions resulted in small differences between the subject's mean relative errors.

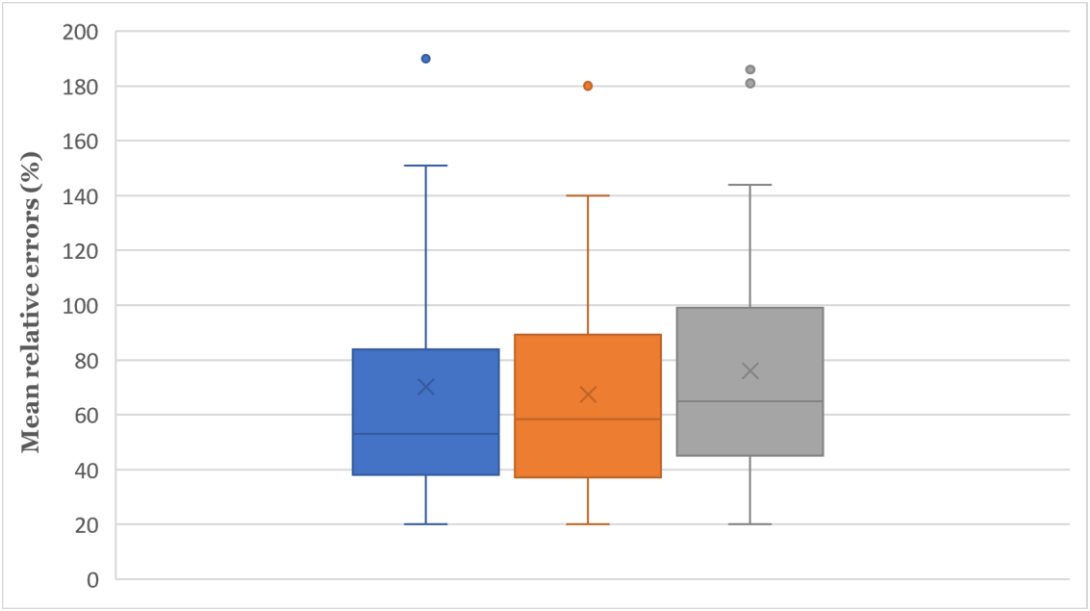


Figure 5: The spread between subject's mean relative errors (%): FOA-tracked (blue), artificial head (orange), FOA-static (grey).

As illustrated in Fig. 5, the three experimental conditions resulted in highly similar sound source distance estimations in the VR environment. But it also shows that there were for all experimental conditions, a large spread between the subject's scores.

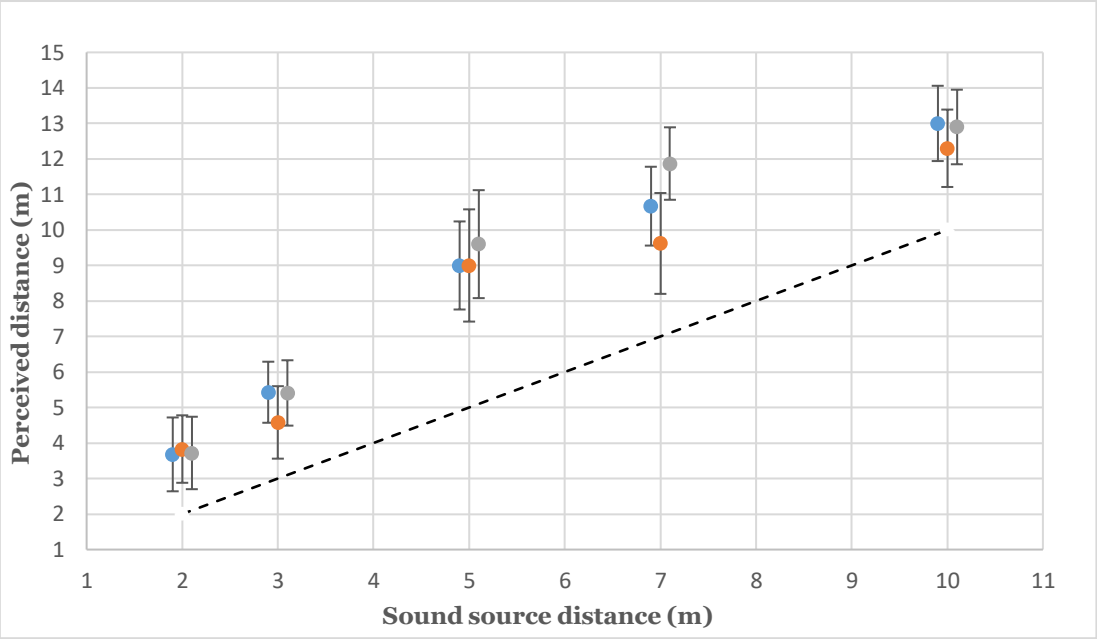


Figure 6: Relationship between the perceived distance (d_{per}) and the sound source distance (d_{src}) (means and 95 % confidence intervals) for each distance level: FOA-tracked (blue), artificial head (orange), FOA-static (grey).

Fig. 6 shows that the subjects clearly overestimated the sound source distances. The black line shows the ideal results.

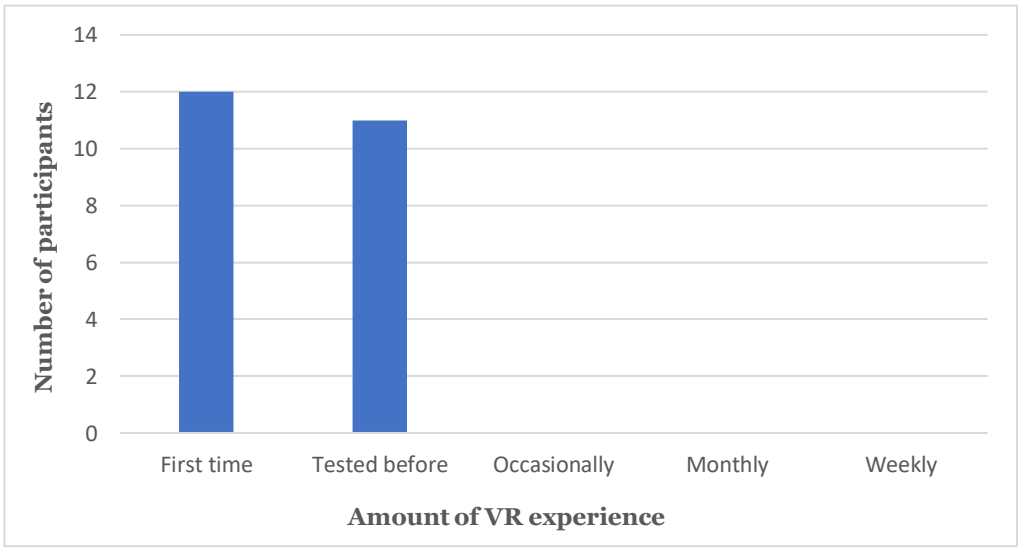


Figure 7: Participants' amount of VR experience.

Fig. 7 illustrates the participant's previous experiences of VR applications. Based on these data, all twenty-three participants in the present study must be considered as novice VR users. No comparison can therefore be made regarding potential differences between novice and experienced VR users.

CHAPTER 5

Discussion

The results presented in the present study indicate that, when a sound source has been recorded with an artificial head or FOA microphone, no significant difference can be shown in how accurately we aurally estimate sound source distances in VR environments. One possible reason might be that the FOA microphone configuration results in capturing enough important reverberation cues in comparison with the artificial head. So even though an artificial head is designed for precisely simulate how we perceive auditory events, it does not result in more accurate auditory distance estimations than the FOA technique.

In the present study, no significant differences were found between the FOA-tracked and FOA-static binaural techniques regarding sound source distance estimations in VR. These results indicate that the binaural cues ILD and ITD were not important for the tasks performed in the experiment and previous studies have shown similar results. In Zahorik, Brungart and Bronkhorst (2005), ILD and ITD are considered doubtful as useful cues for auditory distance perception for faraway sound sources (> 2 m approximately). The results shown in the present study also verifies what Simpson and Stanton (1973) found in their study about that head movements do not help humans estimate the distance to sound sources.

One issue with the comparison made in the present study between the FOA-tracked and FOA-static binaural technique, is that the FOA-static stimuli cannot be considered as 100 % static. Since there was no difference between the FOA-tracked and FOA-static stimuli (rather than the informative widgets), the subjects may have utilized small head movements even when hearing the FOA-static stimuli. In order to have produced 100 % FOA-static stimuli, those stimuli should not have been processed with head tracking. But since no significant differences were found between the artificial head and the FOA-tracked technique, it is still unlikely that a

significant difference would have been found between the FOA-tracked and 100 % FOA-static technique.

5.1 Underestimation of visual targets

In many previous studies regarding auditory distance perception, it has been shown that humans clearly underestimate the distance to faraway sound sources (Zahorik, et al.). However, in the present study the opposite was shown since the participants significantly overestimated the sound source distances. But in contrast to many previous studies, the present study did not only involve auditory stimuli, but visuals as well. Based on this circumstance, the participants' overestimations may have depended on the visual stimuli.

During the pre-study, participants indicated that the virtual hall appeared smaller than the actual virtual measurements of it (the participants did for instance not perceive the back wall as 17 m from the camera position, they perceived it as closer). If the participants in the main study also perceived the virtual hall as smaller than intended, then they likely perceived the virtual loudspeaker as closer than it was.

Imagine a subject hearing one of the sound sources positioned 10 m from the camera position. The subject aurally estimates the sound source distance correctly and positions the loudspeaker at what she perceives as 10 m away. But if the subject perceives the loudspeaker as closer than intended, then she will submit an incorrect score. Because instead of 10 m away, she may have positioned the loudspeaker at 15 m from the camera position. That is, if the participants in the present study perceived the virtual environment as smaller than intended (i.e. the size of the real-world hall), it might be why overestimations of sound source distances were shown in the present study.

Regarding underestimations of visual target distances in VR environments, it has been shown in several previous studies. In an experiment, Thompson, Willemsen, Gooch, Creem-Regehr, Loomis and Beall (2004) showed that subjects clearly underestimated target distances in VR environments when using HMDs. And in Plumert, Kearney, Cremer and Recker (2005) they concluded that our visual distance perception in VR environments may be more accurate when using large-screen displays rather than HMDs. This kind of potential 'compression effects' induced by HMDs probably affected the results shown in the present study.

5.2 Front-back errors

In the results it was shown that one subject perceived all artificial head stimuli as located behind the head. During the listening tests, other subjects also indicated that some stimuli appeared as located behind their heads.

Whether or not those stimuli were the artificial head recordings is unfortunately unknown since the subjects still choose to, by listening to the auditory cues in the recordings, position the loudspeaker at adequate places in front of them (the instructions said that all sound sources were positioned in front of them).

One effect that might have resulted in participants' perceiving sound sources as located behind is the front-back error effect (i.e. when a sound source with frontal incidence is incorrectly perceived as located behind the head or vice versa). While this effect might have affected the scores for the artificial head or FOA-static stimuli, it is very unlikely to have occurred when listening to the FOA-tracked stimuli.

First, if the subject for instance would have turned her head to the left, she would have noticed an apparent SPL difference between the input signals received by her right and left ear. That is, the binaural cue ILD (and probably ITD) would have helped the subject estimate the direction of the sound source. Thus, the subject would probably have realized that the sound source was not positioned behind the listening position. Furthermore, in an experiment by Iwaya, Suzuki and Kimura (2003) it was shown that head movements clearly helped reduce front-back errors. However, based on the notion about IHL it is interesting to speculate on if it resulted in participants' perceiving sound sources as located behind them and whether the artificial head technique may have resulted in more IHL effects than the FOA-static.

The design of the artificial head used in the present study (Neumann KU 100) automatically results in a unique HRTF, while the FOA technique utilizes conversions followed up by implementations of HRTF filters resulting in another HRTF model (Resonance Audio by Google, 2019). While these two HRTF models probably produced slightly different ear input signals, it would have required a dedicated study to find out if either of them generally resulted in more IHL effects than the other. But as mentioned before, IHL can also occur when a listener's ears receive identical or highly similar signals. And due to the radical different designs of the two microphones compared in the present study, one could suspect that the artificial head technique resulted in more similar ear input signals.

When a sound source is in the median plane, both ear input signals can be expected to be roughly the same (Blauert, 1997). That is, sound waves which travel straight to the ears (i.e. direct sound) provide both ears with the same information while the following reflections, depending on the surroundings, input roughly the same signals to the ears. And since an artificial head is designed for simulating such a sound event (i.e. binaural hearing), reproductions of artificial head recordings in the median plane should input highly the same signals to both ears. But the FOA microphone on the other hand, which consists of four cardioid capsules pointing in

different directions may, after the conversions, result in less identical ear input signals and thus, result in less IHL effects.

Finally, the choice of closed-back headphones in the present study can be debated since open-back headphones "…are often thought to have better externalization than closed headphones because of open headphones' lower acoustic impedance at the ear, which provides a more natural sound." (Boren, & Roginska, 2011, p.7). Thus, the choice of headphones may have affected the results negative in producing more IHL or front-back errors than a pair of open-back headphones would have. But in an experiment by Boren and Roginska (2011) closed-back headphones was shown to result in better externalization than two pairs of open-back headphones. In the study, Boren et al. argued for the headphones flatter (i.e. better) frequency response as the reason for the better externalization. Thus, if the headphones used in the present study produced more IHL effects than another pair would have had, the frequency response rather than the type of headphones is more likely to have been the issue.

5.3 Conclusion

In the present thesis, the artificial head, FOA-tracked binaural and FOA-static binaural techniques were compared regarding auditory distance perception in VR. No significant differences were shown between them regarding how accurately subjects estimated sound source distances in a VR environment. In the results, it was shown that the subjects clearly overestimated the sound source distances used in the study. A possible reason is that the subjects underestimated the size of the virtual environment.

5.4 Future work

During the work with the present thesis, some ideas possibly useful for future researches have come to mind.

Firstly, the method utilized in the present study may benefit from a data collection regarding how the involved participants perceive the size of the VR environment. It would probably be beneficial to compare each subjects' auditory distance estimations with their size estimations of the virtual environment. The ability to show *how* affected subjects were by visual stimuli could help increase the validity of these kind of future studies.

Furthermore, regarding the comparison between the artificial head and FOA technique, it would possibly be valuable to implement an option where subjects can choose if they perceive the sound source as located behind their heads. Such an option could help researchers draw conclusions about

whether either of the techniques significantly result in IHL or front-back errors.

Finally, it would be interesting to involve both novice and experienced VR users in future experiments. Based on the notion about familiarity and auditory distance perception, it would be exciting to find out whether that relationship also applies to sound events in VR environments.

References

Begault, D.R. (1994). *3-D sound for virtual reality and multimedia*. Boston: AP Professional.

Blauert, J. (1997). *Spatial hearing: the psychophysics of human sound localization.* (Rev. ed.) Cambridge, Mass.: MIT Press.

Boland, F., Corrigan, D., Gorzel, M., Kearney, G., & Squires, J. (2012). Distance Perception in Virtual Audio-Visual Environments. *Audio Engineering Society Conference Paper, presented at AES 25th UK Conference – Spatial Audio in Today's 3D World.* Retrieved November 21, 2018 from http://www.aes.org.proxy.lib.ltu.se/tmpFiles/elib/20181118/18119.pdf

Boren, B., & Roginska, A. (2011). The Effects of Headphones on Listener HRTF Preference. *Audio Engineering Society Convention Paper 8537, presented at the 131st Convention.* Retrieved April 1, 2019 from http://www.aes.org.proxy.lib.ltu.se/tmpFiles/elib/20190402/16063.pdf

Burdea, G. C., & Coiffet, P. (2003). *Virtual Reality Technology [Electronic resource]*. John Wiley & Sons, Inc. / Engineering.

Epic Games. (2019). Unreal Engine 4 (Version 4.21) [Computer Program]. Available at https://www.unrealengine.com

Everest, F. A. & Pohlmann, K. C. (2015). *Master handbook of acoustics*. (6. ed.) New York: McGraw-Hill.

Gardner, M. B. (1968). Proximity Image Effect in Sound Localization. *Journal of the Acoustical Society of America, 43,* 163. Doi:10.1121/1.1910747

Gardner, M. B. (1969). Distance Estimation of 0° or Apparent 0°-Oriented Speech Signals in Anechoic Space. *Journal of the Acoustical Society of America, 45*, 47-53. Doi:10.1121/1.1911372

Hong, J. Y., Lam, B., Ong, Z-T., Ooi, K., Gan, W-S., Kang, J., Feng, J., & Tan, S-T. (2019). Quality assessment of acoustic environment reproduction methods for cinematic virtual reality in soundscape applications. *Building and Environment, 149,* 1-14. Doi: 10.1016/j.buildenv.2018.12.004.

Hur, Y., Park, Y-C., Lee, S-P., & Young, D. H. (2008). Efficient Individualization of HRTF Using Critical-band Based Spectral Cues Control. *Audio Engineering Society Convention Paper 7447, presented at the 124th Convention.* Retrieved April 3, 2019 from http://www.aes.org.proxy.lib.ltu.se/tmpFiles/elib/20190403/14577.pdf

Iwaya, Y., Suzuki, Y., & Kimura, D. (2003). Effects of head movement on front-back error in sound localization. *Acoustical Science and Technology, 24*(5), 322-324. Doi:10.1250/ast.24.322

Izotope. (2019). Insight (Version 1) [Computer Program]. Version 2 available at https://www.izotope.com/en/products/mix/insight.html

Mershon, D. H., & King, L. E. (1975). Intensity and reverberation as factors in the auditory perception of egocentric distance. *Perception & Psychophysics, 18*(6), 409-415. Retrieved January 15, 2019 from https://link.springer.com/content/pdf/10.3758%2FBF03204113.pdf

Noisemakers. (2019). Ambi Converter [Computer Program]. Available at https://www.noisemakers.fr/ambi-converter/

Paquier, M., Côte, N., Devillers, F., & Koehl, V. (2016). Interaction between auditory and visual perceptions on distance estimations in a virtual environment. *Applied Acoustics, 105,* 186-199. Doi:10.1016/j.apacoust.2015.12.014

Plumert, J. M., Kearney, J. K., Cremer, J. F., & Recker, K. (2005). Distance Perception in Real and Virtual Environments. *ACM Transactions on Applied Perception, 2*(3), 216-233. Doi:10.1145/1077399.1077402

Resonance Audio by Google. (2019). Resonance Audio (BETA Version 1.0) [Computer Program]. Available at https://resonance-audio.github.io/resonance-audio/

Rumsey, F. (2017a). *Spatial Audio (Music technology series) [Electronic resource].* New York: Focal Press.

Rumsey, F. (2017b). Binaural Audio and Virtual Acoustics. *J. Audio Eng. Soc., 65*(6), 524-528. Retrieved January 15, 2019 from http://www.aes.org.proxy.lib.ltu.se/tmpFiles/elib/20190108/18784.pdf

Sherman, W. R., & Craig, A. B. (2003). *Understanding Virtual Reality: Interface, Application and Design*. San Francisco, CA: Morgan Kaufmann.

Simpson, W. E., & Stanton, L. D. (1973). Head Movement Does Not Facilitate Perception of the Distance of a Source of Sound. *American Journal of Psychology, 86*(1), 151-159. Retrieved April 2, 2019 from https://www-jstor-org.proxy.lib.ltu.se/stable/pdf/1421856.pdf?refreqid=excelsior%3A351b0c306d23b425805851152d8a8757

Soundfield. (2019). Surround Zone 2 (Version 1.0.0 64-bit VST) [Computer Program]. Available at https://www.soundfield.com/#/products/surroundzone2

Thompson, W. B., Willemsen, P., Gooch, A. A., Creem-Regehr, S. H., Loomis, J. M., & Beall, A. C. (2004). Does the Quality of the Computer Graphics Matter when Judging Distances in Visually Immersive Environments? *Presence: Teleoperators & Virtual Environments, 13*(5), 560-571. Doi:10.1162/1054746042545292

Thresh, L., Armstrong, C., & Kearney, G. (2017). A Direct Comparison of Localisation Performance When Using First, Third and Fifth Order Ambisonics For Real Loudspeaker And Virtual Loudspeaker Rendering. *Audio Engineering Society Convention Paper, presented at the 143rd Convention.* Retrieved March 5, 2019 from http://www.aes.org.proxy.lib.ltu.se/tmpFiles/elib/20190108/19261.pdf

VB-Audio Software. (2019). Voicemeeter (Version 1.0.6.1) [Computer Program]. Available at https://www.vb-audio.com/Voicemeeter/

Vroomen, J., & De Gelder, B. (2004). Perceptual Effects of Cross-Modal Stimulation: Ventriloquism and the Freezing Phenomenon. In G. A. Calvert, C. Spence, & B. E. Stein (Eds.), *The handbook of multisensory processes* (pp. 141-150). Cambridge, MA, US: MIT Press.

Zahorik, P. (2001). Estimating Sound Source Distance with and without Vision. *Optometry and Vision Science, 78*(5), 270-275. Doi:10.1097/00006324-200105000-00009

Zahorik, P., Brungart, D. S., & Bronkhorst, A. W. (2005). Auditory Distance Perception in Humans: A Summary of Past and Present Research. *Acta Acustica united with Acustica, 91*(3), 409-420. Retrieved April 1, 2019 from http://docserver.ingentaconnect.com.proxy.lib.ltu.se/deliver/connect/dav/16101928/v91n3/s3.pdf?expires=1554124798&id=0000&titleid=75000347&checksum=D317185A73C19A06CE3E205341B8E55D

Appendix 1 – Test instructions (in Swedish)

Instruktioner – Lyssningstest i Virtual Reality

Detta lyssningstest går ut på att bedöma avståndet till ’osynliga’ ljudkällor genom att du ska förflytta en virtuell högtalare till samma position som du upplever att en ljudkälla kommer ifrån. Ljudkällan består av en inspelad monolog på cirka 20 sekunder.

För navigering i spelet kommer du använda uppåt- och nedåtpilen samt knapparna ’A’ och ’Y’ på en Xbox-kontroll. (Högtalaren går också att förflytta med vänster spak). Knappen ’A’ används i menyerna. Knappen ’Y’ använder du för att fastställa högtalarens position för varje enskilt ljudexempel.

Totalt kommer du höra 2 st. övningsexempel + 15 st. ’riktiga’ ljudexempel. Ibland kommer du uppmanas att hålla huvudet stilla och titta rakt på högtalaren. Ibland kommer du uppmuntras att vrida på huvudet för att lättare bedöma avståndet till ljudkällan. Uppmaningarna kommer visas på menyer i spelet.

Om du behöver går det bra att när som helst avbryta testet!

Fråga:

Har du någon känd hörselnedsättning? (vänligen ringa in ditt svar)

Ja **Ja (åtgärdad med hjälpmedel)** **Nej**

Appendix 2 – Photographs of the recording hall

Figure 8: Photograph of the recording hall.

Figure 9: Photograph of the recording hall.