Exploration of First-Order Ambisonics Usage in VR Concert Experiences

By Simone Vinkel

Exploration of First-Order Ambisonics Usage in VR Concert Experiences

- The effect on Presence and Perceptual Quality of Spatial Sound -

Master Thesis

Simone Patricia Vinkel

Aalborg University Copenhagen
Sound and Music Computing

Here you can write something about which tools and software you have used for typesetting the document, running simulations and creating figures. If you do not know what to write, either leave this page blank or have a look at the colophon in some of your books.

Sound and Music Computing
Aalborg University Copenhagen

AALBORG UNIVERSITY
STUDENT REPORT

Title:
Exploration of First-OrderAmbisonics Usage in VR ConcertExperiences

Theme:
Master Thesis

Project Period:
Spring 2017

Participant(s):
Simone Patricia Vinkel

Supervisor(s):
Stefania Serafin

Copies: 1

Page Numbers: 63

Date of Completion:
June 2, 2017

Abstract:

This thesis examines the usage of first-order ambisonics in VR concert experinces rendered binarually through headphones. Two experiments have been conducted with each their individual concert experience. The first experiment (N=27, between group) explore 3 sound experiences in regards to the effect on presence, quality and preference: 1) recorded with an ambisonics microphone (H2n), 2) synthesized ambisonics, 3) mono reference. The results did not show any significant differences, but there was a clear preference towards ambisonics with H2n. The second experiment (N=18, within group) explore the subjective perceptual quality of two different ambisonic microphones (H2n and AMBEO VR), as well as measuring and comparison of presence, and also preference. The results did not show any significant differences on presence, and the perpcetual quality scores were fairly equal. There was a preference towards the H2n experience. Both experiments show successful implementations of ambisonics in VR concerts, that preserve high quality and includes the feeling of "being there".

Lyd og Musik Teknologi
Aalborg Universitet København

AALBORG UNIVERSITET
STUDENTERRAPPORT

Titel:
Udforskning af 1. Grads Ambisonics Anvendelse i VR Koncertoplevelser

Tema:
Kandidatspeciale

Projektperiode:
Forår 2017

Deltager(e):
Simone Patricia Vinkel

Vejleder(e):
Stefania Serafin

Oplagstal: 1

Sidetal: 63

Afleveringsdato:
2. juni 2017

Abstract:

Dette speciale omhandler brugen af 1. grads ambisonics i VR koncertoplevelser afspillet binauralt gennem høretelefoner. To eksperimenter er blevet udført, med hver deres koncertoplevelse. Det første eksperiment (N=27, between group) udforsker 3 lydoplevelser med henblik på deres effekt på presence, kvalitet og præference: 1) optaget med en ambisonics mikrofon (H2n), 2) syntetiseret ambisonics, 3) mono reference. Resultaterne viste ingen signifikante forskelle, men der var en klar præference af ambisonics med H2n. Andet eksperiment (N=18, within group) udforsker den subjektive kvalitetsopfattelse af to forskellige ambisonics mikrofoner (H2n og AMBEO VR), samt måling og sammenligning af presence, og også præference. Resultaterne viste ingen signifikante forskelle på presence, og havde nogenlunde lige scoringer i kvalitetsopfattelse. Der var en præference mod H2n oplevelsen. Begge eksperimenter viser succesfulde implementationer af ambisonics i VR koncerter, der bevarer høj kvalitet og inkluderer følelsen af at "være der".

Contents

Acknowledgement

I want to acknowledge my supervisor Stefania Serafin, who has been helpful, patient and joyful through my whole education - and for being the fastest professor to reply on any questions! Furthermore I want to thank associate professor Rolf Nordahl and teaching associate professor Lars Reng, because they both have been valuable and part of the semester projects leading up to the master thesis. The master thesis could not have been done without help and support from my boyfriend Oliver Lønberg, as well as motivation and practical help from my friends Theis Berthelsen and Claes Nielsen. Lastly I want to pay tribute to my former group members Victor Milesen, Dina Smed and Rasmus Lind. This could not have been done without them, and I want to thank them all for keeping me motivated throughout my whole education. I feel fortunate for our last longing friendships.

Aalborg University Copenhagen, June 2, 2017

Simone Patricia Vinkel
<svinke12@student.aau.dk>

Chapter 1

Motivation

In early 2015, during my bachelor thesis in Medialogy, my interest in creating and exploring virtual reality (VR) experiences arose. Together with 3 fellow students, whom I had been working with for 2 prior semesters, we got funds and permission to acquire a 360 GoPro camera rig on behalf of the Multisensory experience lab at AAU-CPH, which has led to several other study groups creating VR content and research. Our own focus at that time was to study how different camera angles in 360 degree videos influenced the user's perception of 'being there' (presence) at a concert in VR, and how it could be used to optimize the experience. We arranged a live concert scenario together with the Danish band Jack in the Middle at the venue KraftWerket in Valby. The bachelor project thesis was also the first time where I was introduced to binaural audio, as we utilized a Bruel and Kjaer dummy head to record the audio (of the concert) in order to give a more realistic sound representation.

On my first year of the master education Sound and Music Computing (SMC), the 7th semester, I continued to do project work with the mentioned students from Medialogy (mentioned as the group), where we focused more on 3D sound in VR. We were interested in exploring how concerts could be perceived as a social experience in VR, and also how 3D sound would influence perceived realism. We made an 'experimental' holographic video sequence using 15 cameras on a line that was angled towards an anchor point placed in the middle of the band (Jack in the Middle) that was playing in front of a green screen. The videos was then implemented in a Unity scene, where we had modelled and designed a venue inspired by the real venue "KraftWerket". We used object-based 3D audio in Unity and evaluated mono sound against 3D sound in a between-group experiment of 60 participants. Results showed that there were a significant difference in perceived realism, where 3D sound was found to be more realistic, and also a significant difference in perceived dimensionality with 3D sound. The presence of another

person (social value) did not add any significance to the experience.

On the 8th semester of the master the group and I continued in the realm of concert experiences for VR. We had been focused much on learning the limits and applicability of 360 degree video capture, and had worked to some extent with 3D sound and binaural audio for headphones reproduction. For this project we desired to use loudspeaker-based spatial audio, namely a 64-channel Wave Field Synthesis (WFS) system. The aim of the study was to test how the addition of audience noise in WFS around the listener, in a mediated virtual reality concert, would influence the perception of presence and quality of experience. We recorded the band Disarray Son (now known as SILQUE) at the event Teufel Bash 8 in Copenhagen, where we learned to navigate and act in a real-life recording scenarios. Many factors can influence and ruin the capture, i.e. it was an outdoor concert and it was raining, and furthermore we were not in control of e.g. spotlights and their direction, which actually ended up ruining the video capture due to overexposure on some of the cameras. The direct inputs from the instruments into the console mixer was acquired, and used for implementation in WFS. The evaluation was performed with a between-group experiment with and without added spatial audience noise. Our study did not reveal any significant difference of adding spatial audience noise in WFS for a VR concert, neither in regards to the feeling of presence nor the acceptability of quality of experience. Qualitative results indicated that the overexposed video and the point of view (which was perceived as too elevated to most participants) decreased the feeling of presence, but that it did not affect the quality of experience.

On the 9th semester we had the opportunity to do an internship with a company. The Group and I wanted to do an internship together where we could focus on both 360 video and 3D sound capture for VR experiences. We all had a wish and desire in going abroad to California, USA, partly because of the music scene, but also because of the many VR events, meet-ups and companies and business opportunities that might occur networking over there. We had planned the trip for a long time, but some unfortunate situations resulted in no internship directly in the US. However, we overcame this by contacting a Danish 360 video company, Panorama Video, of which the co-founder Janus Heiberg was also the man who first and foremost inspired and helped us learning some core aspects of 360 video capture back in early 2015. It seemed destined to be interns of his company, and we agreed with him to travel to California with the prime goals of exploring branches of 360 video content including musical events, and in particular to explore, capture and implement 3D sound (ambisonics) for 360 productions. A month before our journey began in USA, we also had the opportunity to record DR Big Band in collaboration with Sennheiser, where we were introduced to the Sennheiser "AMBEO

VR" microphone. During the internship I gained more knowledge on 3D audio for headphone-based VR applications, and became certain that my master thesis should revolve around this concept.

The group and I had some interesting discussions whether or not 3D sound was really necessary for musical experiences in VR since commercial users are so used to the standard mono and stereo formats, which are widely used for commercial music. Even most VR music videos today are not utilizing the potential of spatial audio, and so many questions came up to which degree spatial audio can be applied in musical VR applications and still deliver an immersive, pleasant and high quality experience to the user.

This brings me to the very present. After working together with the group for several years and through many good projects, I decided to do the thesis on my own. Group work can be very beneficial, as it teaches one how to work in a team that consists of individuals with different skills and competences. Communication is key, and is another ability that gets perfected. However, it is just as important to be capable of working independently, since it is another valuable skill to have when working in the professional industry.

Chapter 2

Introduction

Virtual Reality (VR) has existed for many decades, but is finally meeting its commercial era. According to Goldman Sachs Research, VR (and Augmented Reality) is assumed to become worth $80 billions by 2025, where Live Events takes $4,1 billions of the total, and video entertainment $3,2 billions [1].

Ambisonics was developed back in the 1970s, and has gained recent popularity for possible commercial usage, and it seems that it is becoming the standard spatial audio rendering method for VR applications and 360 degree videos. The distribution possibilities for first-order ambisonics (FOA) in 360 degree videos has increased particularly within the last two years, where big platforms like YouTube and Facebook both started supporting spatial audio (i.e. ambisonics) playback in 2016 [2], [3]. In addition, YouTube also started supporting live streaming of 360 degree content, with a launch of live streaming the Californian music festival Coachella. Related to music, many artists and bands (e.g. Björk, Gorillaz and Paul McCartney to mention a few) are following the trend and have published 360 degree video concerts and/or music videos, but interestingly a majority of these videos have been rendered with traditional stereo mixes, and have not utilized head-tracked audio that follows the rotation of the video.

Typical consumer audio for headphones is rendered and experienced from one perspective, but as VR and 360 videos permits full spherical view, it can be assumed that the sound also naturally should change according to one's head movements. When mediating true-to-life events such as concerts for VR, it is frequently desired to induce the sense of presence (feeling of 'being there') and create realistic experiences, and maintain high quality and enjoyability. It is often claimed that "immersive technologies" such as spatial audio can bring the experience to the next level. The question is how, or to what degree, and how it affects the perceived quality. This leads to the research question of this thesis:

To what degree does First-Order Ambisonics (via headphones) influence the subjective perception of presence and perceived sound quality in mediated live concert experiences for Virtual Reality?

Chapter 3

Background

This chapter will describe the theoretical and technical background that is relevant to the research question. The first section reviews related work. The second section defines the psychological and perceptual terms of immersion, presence and sound quality, and reviews methods for measuring these variables. The third section describes theory related to human spatial hearing. The fourth section presents the state-of-the-art technologies related to ambisonics 3D sound, VR, 360 degree video rendering and capture.

3.1 Related Work

A review of related work will be presented. It has been difficult to find empirical literature that contains all components of the research question, i.e. perceptual evaluation of presence and sound quality of VR concerts with ambisonics, but much material exists towards separate components. Many studies indicate that spatialized sound (including ambisonics) can induce the subjective feeling of presence [4], [5], [6]. In [7] from 2015 they examine the effect of presence in an auditory virtual environment (VE) with 3D sound reproduction. Using a loudspeaker setup, they compared spatialized versus non-spatialized audio renderings, and measured presence using both psychological and physiological data. Their results showed greater presence ratings for spatialized sound, and they state the there was a correlation between the psychological and physiological data.

In [8] from 2007 they compared perceptual differences in terms of spatial quality and localization between 3 sound reproductions: transaural, ambisonics, and stereo. They used 4 sound scenarios: outdoor, indoor, people talking with background music, and an electric guitar concert. They rated on envelopment, immersion, representation, readability, realism and overall quality. Their results show

that ambisonics provide a good sense of immersion, but not so well localization and readability of the sound scene, compared to the two other methods. Vice versa, transaural and stereo had less immersion but better localization. They conclude that there seems to be a trade-off between immersion and precision. The test was only audio-based and did not include any visuals.

In [9] from 2011 they study subjective comparisons between ambisonic soundfield recordings rendered through stereo and binaural. They use 4 recordings: an orchestra, solo piano, field recording with a duck flying over one's head, and a field recording with footsteps behind listeners. 5 questions were asked for each recording regarding the attributes: wideness, depth, naturalness and presence. 11 subjects were used, and had familiarity with live concert music. Their results show a higher rate of preference towards stereo. They point out that participants rated naturalness and presence in an incoherent way, and that people might prefer what sounds "unnatural" because it is closer to the industry standard, or the "pop music" standard. They did not include any headtracking or visual stimuli.

In [10] from 2010 researchers at BBC investigated in the potential advantages of broadcasting with FOA. They conducted subjective listening tests with 18 subjects, comparing stereo, 5.1 and FOA using loudspeaker setups. 5 different productions was used, 3 music pieces and 2 radio dramas. Participants rated their prerence using the MUSHRA standard listening test with a hidden reference and two hidden anchors. Quantitative results did not show any clear preference, but the qualitative data revealed that for music ensembles ambisonics was preferred over 5.1 and stereo. Their results also indicate that there were not nesecarily any correlation between preference of experience, and the participants describing being present at the performance with ambisonics reproduction.

Very recently, in May 2017, the authors of [11] released the second part of a comparison of ambisonics microphones. They experiment with performances of five commercially available ambisonics microphones, which are used in relation to 360 degree video experiences: Soundfield MKV, Core Sound TetraMic, Sennheiser Ambeo, MH Acoustics Eigenmike, and Zoom H2n. The sounds were reproduced via a loudspeaker array, and the experiment was conducted through listening tests, where timbral quality and localization were measured. They used 21 participants with varying expertises in audio listening. The results showed that the Soundfield MKV and Eigenmike produced significantly better results. Furthermore, they state that the Ambeo and H2n were indistinguishable from each other in regards to high frequency timbre. For low frequency, the Zoom H2n was rated as thin in comparison to the Eigenmike. Ambeo and Eigenmike both had better directionality compared to the others, and they conclude that the Zoom H2n performed

worse overall in comparison to the other microphones, but that it still performs quite well despite its low price.

3.2 Psychology & Perception

This section has two parts concerning the psychological and perceptual variables of the research question. First part will seek to clarify definitions of presence, and how it is distinguished from the term immersion. It will also describe methods for measuring presence. The last part defines spatial sound quality, and how it can be measured.

3.2.1 Immersion & Presence

Virtual Reality serves as a medium that essentially seeks to deliver immersive and strong presence inducing experiences. But what does that really mean? It is important to clarify the meaning of these terms, as there exists numerous definitions of what immersion and presence are. First of all, there exists an inconsistent usage of both terms, and it even appears that they are expressed as meaning the same. This section will go through some of the definitions relevant for this project. First section describes a distinction between immersion and presence, second section goes through the many definitions and variables of presence, and the last section describes how to measure presence.

According to Mel Slater, immersion and presence are strongly related, but are separable terms. Immersion can be defined as an objective measurement that the technology offers. A system or technology is said to be more immersive, the more fidelity and tracking it offers [12]. There can be various levels of immersion (technology), i.e. mono reproduction can be said to be less immersive than surround sound, and 2D screen display less immersive than VR HMD.

Presence on the other hand is a response to the immersive system, a subjective perception of the stimuli through a medium [12]. Slater makes a comparison of the distinction between immersion and presence to the concept of wavelength distribution. A color can be described objectively by its wavelength (like immersion), but the perception of a color can differ, e.g. a color can be perceived as being the same despite different wavelength distributions (like presence) [12].
There are numerous definitions of immersion and its relation to presence, engagement, involvement etc., which might be relevant for game or interaction-based measures. For this study Slater's definition seems to apply, as the level of immer-

sion can be referred to different audio reproduction techniques. The level of immersion can be described as "low", "medium" or "high" in relation to each other. In summary, for this study, immersion is defined as being the level of immersive technology, and presence is the subjective response to the applied technology.

3.2.1.1 Defining Presence

Presence originates from the word "Telepresence", where the word "tele" refers to the senses being transported to another place [13]. The term is used across diverse fields of study such as psychology, philosophy, medicine and engineering, and it makes it difficult to come up with a simple and unifying definition [13]. However, to compare and relate studies of presence, it is critical to define and ensure what is being measured. Marvin Minsky defined (tele)presence in "Omni Magazine" in 1980 as:

> "Telepresence emphasizes the importance of high-quality sensory feedback and suggests future instruments that will feel and work so much like our own hands that we won't notice any significant difference. [...] The biggest challenge to developing telepresence is achieving that sense of 'being there'." [14].

Through the 1990's and early 2000 many researchers came up with their take on what presence is. These includs, but are not limited to, the following definitions:

- "The perceptual illusion of nonmediation" – Lombard and Ditton [15]
- "The subjective experience of being in one place or environment, even when one is physically situated in another"– Witmer and Singer [16]
- "Telepresence, the phenomenal sense of "being there" including automatic responses to spatial cues and the mental models of mediated spaces that create the illusion of place" – Biocca [17]

Lombard and Ditton made a comprehensive literature review in 1997, and conceptualized 6 aspects of presence [15]:

- **Presence as Social Richness:** Is related to the interpersonal communication, the intimacy and immediacy such as eye-contact, conversations, smiles etc.

- **Presence as Realism:** Accurate representation of events and people, that look, sounds or feels real.
- **Presence as Transportation:** Is divided into three sensations of “being there” in a mediated experience: “You are there”, “It is here” and “We are here together”. Also related to the "Suspension of Disbelief".
- **Presence as Immersion:** Can be divided into "Psychological Immersion" and "Perceptual Immersion", where the first is described as involvement, engagement and absorption into the virtual experience and narrative, and the latter is described as the degree to "submerge" the perceptual system of a user.
- **Presence as social actor within medium:** A form of presence that happens when a user try to interact with entities (e.g. a virtual character), and overlook the fact that it is mediated and non-real.
- **Presence as medium as social actor:** A social response to the technology or medium, which is perceived as a social entity by the user.

3.2.1.2 Measuring Presence

The most applied method to measure presence is through self-reported measurements, where questionnaires are often used. Ideally it should be measured while in the experience and feeling presence, but paradoxically that would ruin the illusion of non-mediation. An approach is to ask the participant to fill out the questionnaire immediately after the experience, while the memory of the experience is fresh in their mind. Others suggest to use physiological measures to correlate with the self-reported measures, but the opinions about this is divided among the scholars. In "Towards a Robust Quantitative Measure for Presence" [18] the authors states that presence is a subjective sensation, and therefore objective measures, such as heart rate, cannot be correlated with presence.
In regards to self-reported questionnaires, there have been many attempts to create valid and reliable questions that actually measure presence, e.g. Lombard & Ditton Questionnare [19], Slater-Usoh-Steed Questionnaire (SUS) [20], or the ITC-Sence of Presence Inventory (ITC-SOPI) [21]. The Lombard and Ditton questionnaire is based on the aforementioned presence conceptualizations, and its full length consists of 103 items, that are related to presence, and also the overall perceived quality of video and sound, prior experience of the media, and more. The SUS questionnaire is very short, and consists only of 6 items related to presence, rated on a Likert scale of 1-7. It relates to three aspects of presence: Sense of being there, extent to which the experience becomes more real than reality, and the extent to which the experience is thought of as a place visited. The ITC-SOPI questionnaire was developed to provide a comprehensive, reliable and valid approach to measure presence,

based on previous attempts and criteria. The ITC-SOPI questionnaire was tested on more than 600 people, and revealed 4 factors contributing to presence: Sense of Physical Space, Engagement, Ecological Validity, and Negative Effects.The full revised questionnaire consists of 44 items, where scores from questions related to each of 4 factors are summed and their means calculated. Questions are asked on a 5-point Likert scale with the extremes "Strongly Disagree" to "Strongly Agree". Interestingly, all three questionnaires were developed at the same time period around the year of 2000. All of them, among many others, have been widely used, and are acknowledged by the scientific community.

3.2.2 Spatial Sound Quality

Sound quality can be defined in terms of its physical properties, such as noise, frequency response, etc. Perceptual sound quality however is based on judgements from listeners, being experts/trained or naïve listeners [22]. Judgements are then quantified, either through perceptual measurement (strength of individual auditory attributes), or affective measurement (overall impression) [22]. There are numerous of standards, or recommendations, for assesing perceptual sound quality. Many standards treats spatial quality as lower level attributes, e.g. one or few entities that are part of one single rating of the overall sound quality [23]. With the emergence of spatial sound (including binaural playback, surround sound etc) in consumer applications, it becomes more important to specify attributes related to spatial qualities [23]. Most consumer audio is often created with an unnatural and artistic purpose, and does not seek to create realistic reproductions, whereas the aim of spatial audio reproduction is often to create a believable "being there" sensation. However spatial sound can also be non-realistic for creative purposes, being "hyperreal" [23], e.g. a concert where the high quality audio tracks for each instrument has been positioned at their visual appearance rather than coming from the PA speakers. Different versions of spatial sound reproduction should be comparable, by defining relevant quality attributes and compare their magnitude differences. One of the most used attributes assesed in spatial audio quality (especially for loudspeaker setups) is localization. Interestingly, studies have shown that there is a low correlation between accurate localization and listeners' preference [24]. In addition to quality attributes preference should also be assessed.
For the actual spatial quality attributes, there have been attempts to define these [23], [22], [25], [26] . The Spatial Audio Quality Inventory (SAQI) has developed an inventory defining the perceptual (spatial) quality, with the purpose of *"assessments of unimodal or supramodal auditory differences between technically generated acoustic environments (VAES) as well as with respect to a presented or imagined acoustic reality."* [25]. It uses scale ratings with labels as extremes, e.g. "lower-higher" or "less-more". In total there are 48 items describing perceptual qualities, with 8 overall categories

realting to: Timbre, Tonalness, Geometry, Room, Time behaviour, Dynamics, Artifacts, and General. Even though it was not developed with the intention of testing naïve listeners, then it still can be used to do so as long as the experimenter (me) construct more intuitive formulations in order to make them understandable for a non-expert audience [25]. The Internationanl Telecommunication Union (ITU) has many recommendations for sound quality assesment, e.g. ITU-R BS.1284-1 *"General methods for the subjective assessment of sound quality"* [26]. This is more oriented towards basic audio, but have a few attributes related to spatial audio. They have 7 main attributes being: spatial impression, stereo impression, transparency, sound balance, timbre, freedom from noise and distortions, and main impression. The rating scales goes from 1-5, where 1 = bad, 2=poor, 3 = fair, 4 = good, 5 = excellent. The data should be treated statistically by deriving the mean values and confidence intervals [26].

3.3 Spatial Hearing

Understanding core aspects of spatial hearing, and how the human perceive three-dimensional sound, is important as it relates directly to how 3D audio (for headphones) is implemented. When sound sources reaches the two ears, the brain has to compare the incoming sequence, which is dependant on several cues [27]. This subsection will list and review some of the most important cues for human perception of spatial hearing.

3.3.1 Interaural Time Difference

For horizontal (azimuth) localization, there are two primary mechanisms or cues that play a significant role. One of them is the Interaural Time Difference (ITD), which is most sensitive to low frequency content (below 1.6 kHz) [28]. When a sound source that does not come from directly in front of the listener (0 degrees), reaches to two ears, there will be a differences in arrival time. The ear that is furthest away will receive a delayed version, and there will be a phase difference of the sound source [29], [27].

3.3.2 Interaural Level Difference

The other primary cue for localization in the horizontal plane is the Interaural Level Difference (ILD), and is most sensitive to higher frequencies (above 1.6 kHz) [28]. The head will act as a filter and shadow or attenuate the intensity level of the sound source, so that the ear that is closest to the sound source will perceive a higher sound level. [29], [27].

3.3.3 Cone of Confusion and Head Movements

ITD and ILD are only useful in the horizontal plane, and results in a cone of confusion, where the sound source lies on the surface of a cone. This results in ambiguity of sound localization, e.g. front-back confusionn. However, even just slight movements of the head can solve the cone of confusion, as well as spectral cues [28].

3.3.4 Pinnae and Spectral Cues

For sound source localization in the vertical plane the external ear, called the pinna, is important. It behaves as a linear filter, and alters the frequency spectrum of a sound sources due to reflections and resonances, before reaching the ear canal. The shape of the pinnae is very irregular and varies from person to person. The head and shoulders also reflects and absorbs frequencies which alters the perceived sound [28]. All of these cues results in unique head-related transfer functions (HRTF) for each ear [29], which will be further explained in a subsection 3.4.2.

3.3.5 Cocktail Party Effect

When multiple sound sources are present at the same time, (most) humans are able to focus, to some degree, on selective parts of the sounds, and filter out "noise". This is known as the phenomena "Cocktail Party Effect" [30].

3.3.6 Vision and Ventriloquism

Humans rely to a large degree on vision, and it is arguably the most dominant sense. When visual and auditory stimuli are presented simultaneously, vision often takes over; what is seen is also what is heard. An example is the illusion known as the McGurk effect found in speech perception [31]. Movements of the lips can manipulate the brain into believing the sounds are coherent with what is being seen, even though it is not what is being pronounced.
Another illusion affects the perceived position of sound, and is known as the "ventriloquism" effect. A ventriloquist uses a technique to produce speak without moving the lips, and makes coherent mouth movements with a puppet. Because the visual stimuli appears coherent and synchronized with what is heard, and it is associated with the sound source, the brain interprets the localization of the sound source as being true, even though it actually comes from another point in space [32].

3.4 Technology & SOTA

This section of the background chapter will introduce the reader to the technologies and state-of-the-art methods that are found relevant for this thesis. The first part presents different sound formats, and methods of sound capture and headphone reproduction, the second part defines virtual reality and 360 degree videos, and last the section presents relevant video capture and reproduction methods for VR.

3.4.1 Mono/Stereo

Mono (short for monophonic) sound is the most simple form of audio reproduction that only has one channel, whereas stereo (short for stereophonic) has two channels. A mono source can be played in stereo output, so that the same source is played in both speakers in e.g. headphones, but the source can also be panned between left and right speaker. The word panning originates from "panorama" which means a view of a wide area. Several mono sources can be used to create a stereo mix, which is often done in the music industry. E.g. the guitar is panned to the right, drums in the middle, and bass to the left. A stereo mix can also be downmixed to a monoput. When a stereo sound source is center-panned it is essentially playing in mono, because no left/right panning has been done [33]. Stereo sound does apply some degree of spatiality, but is limited to the frontal horizontal plane, and relies on ILD and ITD cues.

3.4.2 Binaural Audio and HRTFs

Binaural audio includes the spectral cues and pinnae information that humans use to determine sound locations in the vertical plane, i.e. the HRTFs. A small microphone is placed inside each of a person or a dummy-head, and the incoming sounds are affected by the head, torso and pinnae that arrive at each ear before being recorded and stored. The recorded input results in pairs of head-related impulse responses (HRIR), one for each ear, at each specific point in space. The HRTF is then the fourier transform of the HRIR, and is the specific frequency response of a point in space to the point in the ear. Binaural audio can be synthesized with any sound, by convolving it with the HRIR that corresponds to that position in space [34]. However, as each human have varying and different shapes of ears (head, and torso), it produces problems when generalizing HRTFs. Optimally to obtain the most accurate representation of a sound field with best localization abilities, individual HRTFs should be used. However, it is a cumbersome and time-expensive, not to mention unpractical, procedure, and not really possible for the end-consumer to achieve at the moment. However, much research has been done

to find typical characteristics and features of the ears to create generalized HRTFs which actually works for most people [29]. There exists several genereic HRTF databases that are free to use, however many consumer applications that enable binaural audio do not have the option to switch between any HRTFs; as time goes, this will probably change, so that people with their own measured HRTFs can apply these etc. Besides the complications with using non-individualized HRTFs, other problems can be encountered with binaural audio. Typical binaural stereo reproductions miss important aspects such as head movements (by head tracking) and visual cues [29]. However, VR does enable both head tracking and visual cues, which makes binaural audio very suitable.

3.4.3 Ambisonics

This subsection will explain the principle about ambisonics, and how it can be recorded through microphones and/or be synthesized, and also how it is reproduced through headphones.

3.4.3.1 Principle

Ambisonics is a technique used to represent a sound field, rather than being channel-based. It was developed in the 1970s by (among others) Michael Gerzon, and is based on spherical harmonic decomposition [35]. Figure 3.1 shows the first 16 spherical harmonics. The order determines the spatial resolution, and number of channels required [36]. For FOA, 4 components are required. The first is the 0th component that corresponds to omnidirectional pressure, and is denoted **W**. The remaining three components are the 1st order components and corresponds to orthogonal figure-of-eight patterns that each represent directions: **X** is front-back, **Y** is left-right, and **Z** is up-down [29]. 1st order is only based on 4 channels (W,X,Y,Z), but increasing the order will quickly result in a demanding requirement of channels, e.g. 3rd order is based on 16 spherical harmonics, as it is seen on figure 3.1. Another term for FOA is the B-format.

3.4.3.2 Channel Ordering

There are two standards for B-formats; Furse-Maklham(FuMa) [37], and ambiX [38]. They differ in channel ordering, normalization and weighting of orders. FuMa is based on the traditional ambisonics channel ordering, and as such the FOA channel ordering is **W, X, Y, Z**. The newer standard, which Google / YouTube have adopted, ambiX has the channel ordering **W, Y, Z, X**. Furthermore, Facebook uses a non-standard ambisonics format called TBE (short for Two Big Ears, an audio company they acquired in 2016), which is an 8-channel format [39], [40]. Conversions can be made between the different formats, but the fact is that they are all used in different applications, so at the time being one has to create several

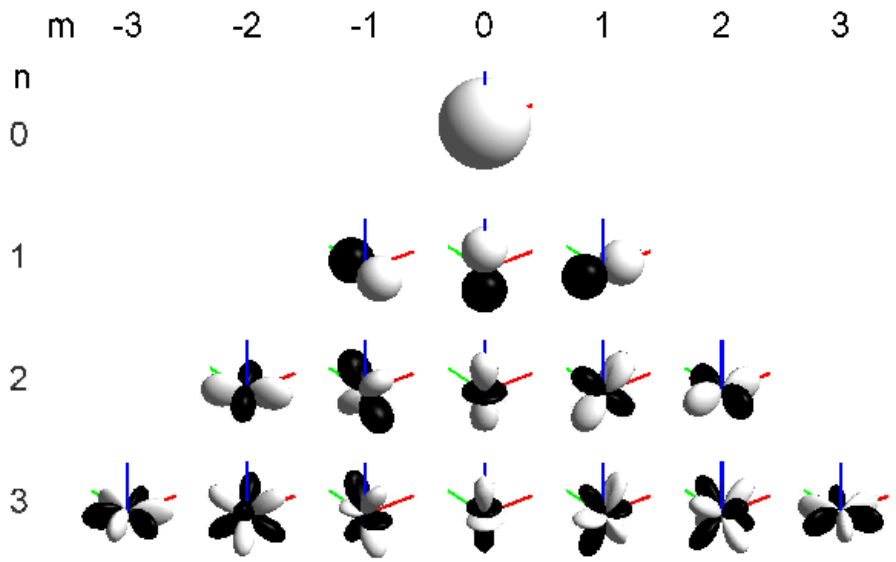

Figure 3.1: Spherical harmonics up to 3rd order, where m is degrees and n is order [36]

mixes that fits to the desired distribution platforms.

3.4.3.3 FOA Microphones

The standard FOA microphone as introduced by Gerzon [35] is based on 4 coincident capsules in a tetrahedral structure, as seen on figure 3.2. The 4 capsules capture sounds in each direction; left-front (LF), right-front (RF), left-back (LB) and right-back (RB). The A-format signal can then be converted to the B-format using simple mathematical principles [29]:

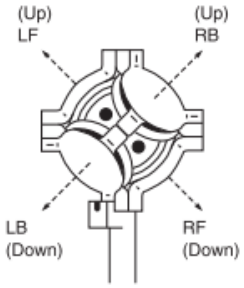

Figure 3.2: Typical tetrahedral microphone setup for capturing a soundfield [41]

W = 0.5(LF + LB + RF + RB)
X = 0.5((LF - LB) + (RF - RB))
Y = 0.5((LF - RB) - (RF - LB))
Z = 0.5((LF - LB) + (RB - RF))

Sennheiser AMBEO VR Microphone
of 2017 there are a variety of FOA microphones to choose from, which ranges in both price and quality. Among the most recent on the market is the Sennheiser AMBEO VR microphone that was released late 2016. Its approximate price-level, at the day of writing, is 14.000 DKK [42]. It has four matched KE 14 capsules in the aforementioned tetrahedral structure, as seen on figure 3.3, and they require phantom power to work. It is capable of full 360 degree spherical recordings. The AMBEO VR microphone comes with a special DIN12M to 4 XLRM split cable, so the microphone can be connected to a recorder with 4 XLR inputs that stores the sounds on 4 tracks. The cable is approximately 0.5 meter short, a very unpractical length, so extension cables are necessary in most cases. It also comes with a Rycote suspension mount, so it can be attached to a microphone stand. Figure 3.3 shows the microphone and assecories, and table 3.1 lists some of the technical specifications.

Sennheiser provide their own software to convert the A-format to B-format, which can be downloaded and used in most DAWs. Their conversion is based on the following calculations:
W = FLU + FRD + BLD + BRU
X = FLU + FRD – BLD – BRU
Y = FLU – FRD + BLD – BRU
Z = FLU – FRD – BLD + BRU

Figure 3.4 shows the interface of the AtoB conversion software. It has few filter correction options, and one can choose the position the microphone was recording

Figure 3.3: The Sennheiser AMBEO VR microphone, cable, and mount. Left side shows the tetrahedral setup [42].

Table 3.1: AMBEO VR specifications

Dimensions	**215mm (H) x 25mm (W) x 49mm (D)**
Weight	**1.060 kg**
Frequency Range	**20 Hz to 20 kHz**
Max SPL	**130 dB(A) for 1 kHz**
Signal-to-Noise Ratio	**18 dB (A weighted)**
Pick-up pattern	**4x cardioid, in A-format arrangement**
Max Recording Quality	**Depends on recorder used)**
Microphone connector	**DIN12M, use enclosed adapter cable to convert to 4x XLR3M**
Power supply	**4x phantom powering (P48)**
Mixing to B-format	**AtoB Converter**

in. The sound image can also be rotated. Lastly is the output format which can be set to either FuMa or ambiX (section 3.4.3.2 will explain the difference).

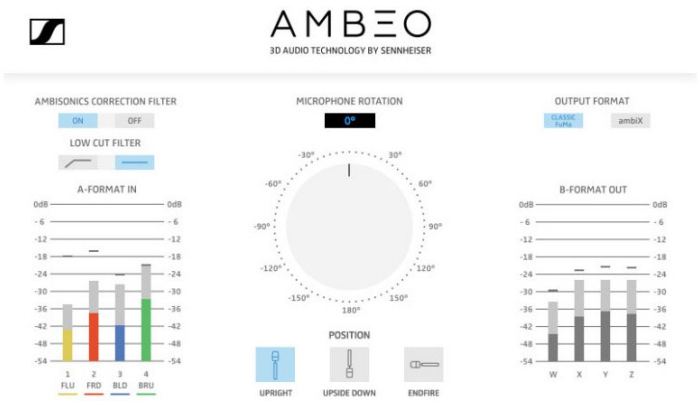

Figure 3.4: Sennheiser's AtoB conversion software [42]

Zoom F4 Recorder

As mentioned, the AMBEO VR microphone requires a recorder, or an audio interface, to record and store the sound data. The recorder should have 4 XLR inputs, and be able to provide Phantom Power to all inputs. Furthermore, an important function is to link the gain controls such that one knob controls the input gain for all inputs. If there is just a slight difference in the input levels for the 4 microphones, the ITD and ILD information are not accurately represented, and it ruins the whole concept of recording the soundfield. A good candidate for these requirements is the Zoom F4 MultiTrack Field Recorder, see figure 3.5. It has the function called "Trim Link", which makes it possible to adjust multiple track input levels at the same time. It has 4 inputs, and can record up to 192kHz 24 bit WAV files.

Figure 3.5: Zoom F4 MutliTrack Field Recorder [43]

Zoom H2n Field Recorder

At the lower end of the price spectrum of FOA microphones, is the Zoom H2n Handy Field Recorder. At the time of writing, it is priced at a humble 1.249 DKK. Originally, the Zoom H2n was not designed to record ambisonics, but has a 4-channel surround sound mode that uses its two sets of build-in microphones: A mid-side pattern in front direction, and a X/Y stereo pattern in rear direction. With a new firmware update 2.0 it can now convert the 4-channel mode into B-format. However, due to the microphone setup, it is only capable of recording horizontal ambisonics, which means that the Z-channel that corresponds to up-down (elevation) will be empty. The AtoB conversion is done internally in the microphone. Figure 3.6 shows a picture of the Zoom H2n and its polar patterns, and table 3.2 shows some technical specifications.

Figure 3.6: Zoom H2n Recorder [44]

Table 3.2: Zoom H2n specification

Dimensions	**113.85mm (H) x 67.6mm (W) x 42.7mm (D)**
Weight	**130 g (without batteries)**
Frequency Range	**50 Hz to 20 kHz**
Max SPL	**120 dB spl (directional), 122 dB spl (bi-directional)**
Pick-up pattern	**Mid-Side + XY**
Max Recording Quality	**48kHz 24 bits (Spatial Mode)**
Power supply	**Two AA batteries, AD-17 USB to AC adapter (DC 5V 1A)**
Mixing to B-format	**Internally in Hardware**

3.4.3.4 Synthesized/Panned Ambisonics

Ambisonic soundfields can also be created artificially, where mono sound sources are positioned and panned using specific software or plugins. To position a mono signal in a three dimensional space encoding equations can be applied [37]:
X = input signal * Cos A * Cos B
Y = input signal * Sin A * Cos B
Z = input signal * Sin B
W = input signal * 0.707

Where A is the angle, and B is the elevation. The multiplication of 0.707 in the W channel should give a more even distribution of levels across the four channels, although 1 is used in some other cases [37], [29].
There are a great amount of plug-ins available for ambisonics encoding to choose from. One of them is the Ambisonics Toolkit (ATK), a free and open-sourced plugin made for the Reaper DAW [45]. It comes with a variety of functions, such as the "PlaneWave" encoder (see figure 3.7), where a mono signal can be placed in space according to azimuth and elevation directions. Using the "RotateTiltTumble" (see figure 3.7) function one can rotate the soundfield, which can be very useful when synchronizing and positioning sounds together with a 360 degree video.

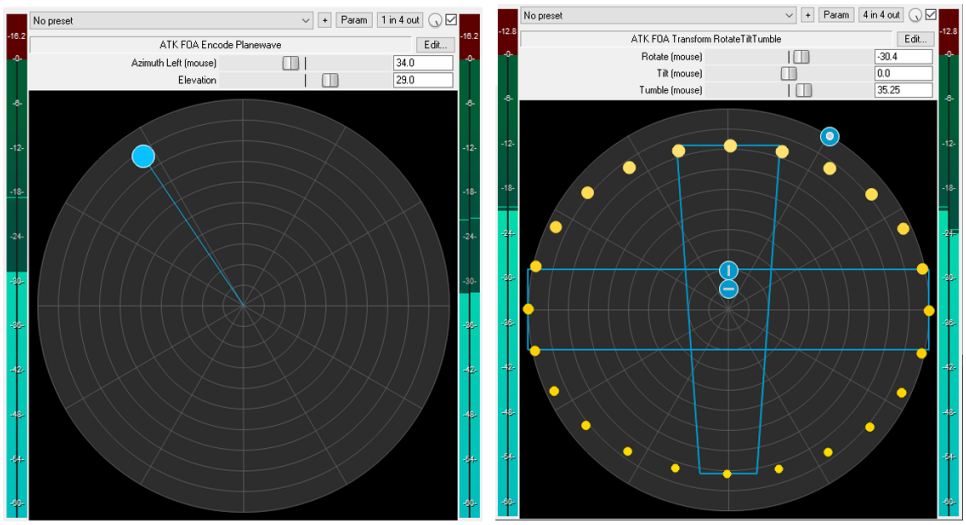

Figure 3.7: Screenshots of the two functions PlaneWave and RotateTiltTumble functions in ATK

Another option is Facebook Spatial Workstation 360, which is also free to use, although not open-sourced. It provides additional features that are really useful for creating ambisonics for 360 degree videos, such as the FB360 Encoder. It is a standalone application that can mux (combine audio and video streams) together 360 videos and different kinds of ambisonic mixes of different orders. There are 6

output formats to choose from, which includes Facebook and YouTube that each has their own requirements for spatial audio in 360 degree videos.

3.4.3.5 Binaural Decoding

The ambisonic B-format are speaker-independent and can easily be decoded into any kind of speaker setup, whether it is mono, stereo, surround sound or binaural. The latter is what is most interesting for VR applications, as the medium primarily, at least for consumers, is used with headphones. A typical binaural decoder uses a virtual loudspeaker array, where pairs of HRTFs are convolved with a loudspeaker that corresponds to the same position. Combining this with head-tracking, the soundfield can also be rotated by applying a rotation-matrix to the signals [46].

3.4.4 Virtual Reality and 360 Degree Videos

Virtual Reality can be defined in several ways, and no clear consensus exists on the term. Some definitions rely solely on the technology, such as Burdea and Coiffet:

"A high-end user interface that involves real-time simulation and interactions through multiple sensorial channels. A VR user is immersed into a computer-generated world via boom or head-mounted display and may "fly" or "walk" through the virtual world and interact with it" [47]

Other definitions are more broad and to some degree philosophical, and focus on the experience of presence, like Steur:

"A virtual reality is defined as a real or simulated environment in which a perceiver experiences telepresence" [48]

This definition does not rely on the applied technology, and could also include mediums such as television. Sherman and Craig determines that there are 4 key elements of VR experiences: a virtual world, immersion, sensory feedback, and interactivity, and comes up with a definition on VR based on these elements:

"Virtual reality: a medium composed of interactive computer simulations that sense the participant's position and actions and replace or augment the feedback to one or more senses, giving the feeling of being mentally immersed or present in the simulation (a virtual world)" [49]

The technologies used to achieve this is first and foremost the use of a head-mounted display (HMD), that enclose the vision of the outside (real world), and

displays the simulated environment. It also provides tracking of a user's position, and HMDs extend the tracking for "room-scale" tracking which allows a limited space for movement. Most often VR experiences are multi-sensorial, and includes audio (via headphones or loudspeakers), or touch (tactile feedback, i.e. vibrations).

One way to simulate or mediate experiences is by capturing spherical videos of real-world scenarios, also known as 360 degree videos. In short, this can be done using multiple cameras that record in every direction from one point, with overlapping field of views. The recordings are then stitched (merged) together in post-production using dedicated software that creates what is called an equirectangular video that can be displayed on special VR players on e.g. YouTube and in some HMDs. 360 Video technology can be used to capture real-life events such as concerts, and the next section will examine two camera rigs that are considered for this purpose.

3.4.5 360 Cameras & Rigs

During my bachelor thesis my study group and I utilized a camera rig consisting of 14 GoPro HERO4 Black cameras, which can be seen on figure 3.8. At that time there were not so many options to choose from in regards to specific cameras for 360 videos, but since then it has expanded tremendously, and many companies now provides 360 cameras ranging in size, price and quality. This section will through the three camera rigs that I find relevant, and I will list the advantageous and disadvantageous of them all. The main advantage of the 14 camera GoPro rig is that it is capable of capturing stereoscopic video, where pairs of cameras are placed in a spherical rig. However, the extreme amount of data gained from 14 cameras recording in high quality is not only cumbersome to deal with, but also fairly irrelevant, unless one needs to capture stereoscopic video material. Another disadvantage is that the rig needs to be placed minimum 2 meters away from objects to avoid "parallax issues", which occurs when the view of each camera has different points. This is not very convenient for places like a small venue, or if the rig is placed in the middle of the audience at a concert. On 8th semester my group and I built a rig made of wood that could contain 7 GoPros, and fully capable of monoscopic 360 video capture. We also made holes for charging cables, which is actually very necessary. The rig can be seen on figure 3.9. Besides the obvious advantage of half the data, using only 7 GoPros, is the fact that the cameras are placed closer to each other, and parallax can be almost avoided when objects are placed minimum 1 meter away, compared to the 2 meters required for the 14 rig. A third option is to use one of the compact and portable devices, such as the Samsung Gear 360 (see figure **??**). It consists of 2 180 degree wide-angle lenses, and is the size of an old webcam or tennis ball. The advantage of this type of camera is without doubt the ease-of-use, portability, less data, and because there are only

2 cameras, there are less problems of parallax errors. During my internship in USA, this camera was used for all recordings made, and we experienced particular problems in low-light conditions, where the ISO could not make up for it without destroying the quality too much. For daylight situations, the camera performed sufficiently well. Figure 3.10 shows the Samsung Gear 360 camera.

Figure 3.8: GoPro 14 Rig

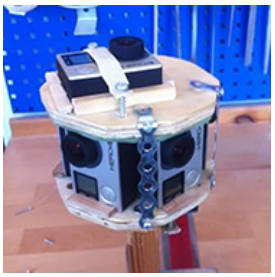

Figure 3.9: GoPro 7 Rig

Figure 3.10: Samsung Gear 360

Both the GoPro-based rigs and the Gear 360 (and other exisitng 360 cameras), have the problem of overheating and poor battery performance. The cameras are also very sensitive to temperature, where a hot day in the sun could drain them out within 20 minutes, but a cold winter day they could possibly last for 2 hours, which was the case when recording a Christmas Parade with the Gear 360. This is especially an important aspect when recording with the GoPro rigs, because if one camera dies, it ruins the whole capture. Therefore it is crucial to charge the cameras while recording to secure as much recording time as possible for all cameras. This of course involves many cables, chargers and power outlets, but needs to be considered and prepared before recording. A table of the specifications of the GoPro 7 rig and the Samsung Gear 360 is listed below:

	GoPro 7 Rig
+	Overall better video quality
+	Long battery life while charging
-	Large amount of data
-	1 meter distance to avoid parallax
-	Many cameras to opreate

(a) GoPro 7 Rig

	Samsung Gear 360
+	Adequate amount of data
+	Easy and quick stitching
+	Easy to operate
+	Live preview
-	Poor quality in low light conditions
-	Need phone to change ISO

(b) Samsung Gear 360

Table 3.3: Advantages and Disadvantages

3.4.6 Head-Mounted Displays

Among the most popular HMDs on the consumer market are the HTC Vive, Oculus Rift CV1, Gear VR, and at this very moment Microsoft are soon to release development kits for their version of VR HMD. The requirements of a HMD for this thesis is its capability of playing 360 degree videos, that it supports binaural playback of ambisonics audio, and of course headtracking. The HTC Vive and Oculus CV1 both needs powerful PCs to run, and are wired. The Oculus Rift CV1 has in contrast to all other HMDs integrated headphones. As neither Oculus CV1 nor HTC Vive have any native 360 video player capable of playing 360 videos with ambisonics, there are ways to work around this, e.g. by the using the Ozo Preview desktop video player, and simply preview the video in VR screen display mode. The Samsung Gear VR is another type of HMD and it differs in way that it runs on a Samsung smartphone, and is therefore wireless. It is also lighter to wear on the head (318 gram) compared to HTC Vive (555 gram) and CV1 (380 gram). The Gear VR has two 360 players capable of ambisonics playback: Oculus Video player and Samsung VR. It should be mentioned that the CV1 also has the Oculus Video player, but that it is not yet able to playback spatial audio on that device. The main disadvantage of using a Gear VR is that it has no monitoring, so only the person wearing the HMD knows what is going on. There are few apps, e.g. "AirDroid" that mirrors the phone's screen to a PC, but it is very slow and there is much lag, and it will quickly drain the phone's battery.

Figure 3.11: HTC Vive [50]

Figure 3.12: Oculus Rift CV1 [51]

Figure 3.13: Samsung Gear VR [52]

3.5 Summary and Goal

With the knowledge gained from the background chapter, the next step is to apply this knowledge to conduct experiments that seek to answer the research question.Different assesments of measuring presence and quality will be explored, as wellas different implementations of first-order ambisonics in VR concert experienc

Chapter 4

Experiments

This chapter describes the procedure of the two experiments that were conducted to explore the research question. It is divided into two main parts that will explain the goal, production, methods, test setup, results and findings of each experiment.

4.1 Experiment 1

- Preliminary Exploration of FOA in headphone-based VR Concerts and the effect on Presence, Overall Quality and Preference

The first experiment conducted was meant as a probe to investigate how a VR concert with sound implementations of recordings from an FOA microphone, a synthesized FOA mix, and a mono mix respectively would affect presence, the overall quality and also to find out what was the preferred listening experience of the participant(s). The experiment was divided into two parts; a quantitative between-group experiment, and qualitative interviews.

4.1.1 Goal

The goal for this experiment was to evaluate how the implementation of FOA in a VR concert experience affected presence and the overall quality. Two kinds of ambisonics implementations were investigated: FOA recorded with a Zoom H2n, and FOA synthesized and designed using the ATK toolkit with audio tracks from a console mixer. A mono mix using the same audio tracks as mentioned above was also created, to use as a baseline to test against the ambisonics mixes.
Lastly the goal was also to investigate which of these three sound representations were preferred when asking the participants.

The requirements for the final productions were to have 3 identical 360 degree videos of a concert each with a different sound representation: 2 kinds of ambisonics, 1 mono. The HMD chosen for the experiment was the Samsung Gear VR using Oculus Video player.

4.1.2 VR Production

The production was made using materials gathered during my internship in the United States on the 9th semester. The concert was recorded at the legendary blues venue Antone's in Austin, Texas on Christmas Eve the 25th of December 2016, and featured an upcoming female blues guitarist Jackie Venson.

4.1.2.1 Preparations

A couple of days before the actual recording day, some tests regarding video and audio recordings were done at Antone's to ensure the best possible capture. The camera that was utilized for this concert was the Samsung Gear 360, and as mentioned in section 3.4.5 it does not perform well under low-light conditions and especially not with high ISO settings. Therefore it was necessary to test how low the ISO could be set to in the specific light conditions, as to not end up with a completely black image. The conclusion was that to get the clearest image without too much noise, it could go as low as 400 ISO (lowest possible setting), as long as there were no extreme light changes during the concert. The reason that Gear 360 was used is simply because this was the camera that I was experimenting with in USA, and the only camera that I had access to at that time as well.
Audio-wise, it was important to ensure that all channels were being recorded from the mixer to my computer. This meant that I had to coordinate with the sound guy at Antone's, and I also had to download specific drivers such that the computer would be able to connect with the mixer. The digital mixer at Antone's was a DiGiCo S21, and the sampling rate was 96kHz. The computer and mixer was connected using an AB USB cable, and was recorded directly to the DAW Reaper in the same sample rate as the mixer used. The incoming channels were labeled identical as the sound guy had named them on the mixer. A test recording was done on the 21st of December, and the Zoom H2n recorder was also tested to get an idea of how the gain settings should be set to avoid clipping and distortion.

Here is a list of equipment and things that were needed for the recording:

- Fully charged Samsung Gear 360 camera
- Empty micro SD Card
- Samsung Galaxy Smartphone (to change ISO)

- Zoom H2n recorder
- Computer (laptop) with Reaper and installed DiGiCo drivers
- Tripod
- Ball head clamp (to attach Zoom H2n on the tripod)

4.1.2.2 Recording

The recording took place at Antone's Nightclub in Austin Texas on Christmas Eve the 25h of December 2016. Arriving early evening at Antone's ensured enough time to set up the equipment, before miss Venson was playing. An extra Samsung Gear 360 camera was brought, in case the first one should overheat or run out of battery too fast. The Zoom H2n ran on AA batteries. The camera was set to ISO: 400, Resolution: 4K (3840 x 1920), Framerate: 30 FPS, White Balance: Auto. The Zoom H2n was set to spatial audio mode, and at its maximum quality of 48 kHz/24 bits. The audio tracks recording on the computer from the mixer were recorded at 96kHz/24 bit, because it had to be identical to the mixer's settings. In total nearly 2 hours of the concert were recorded, where the camera was switched approximately an hour in. An overview of the recording setup can be seen in figure 4.1.

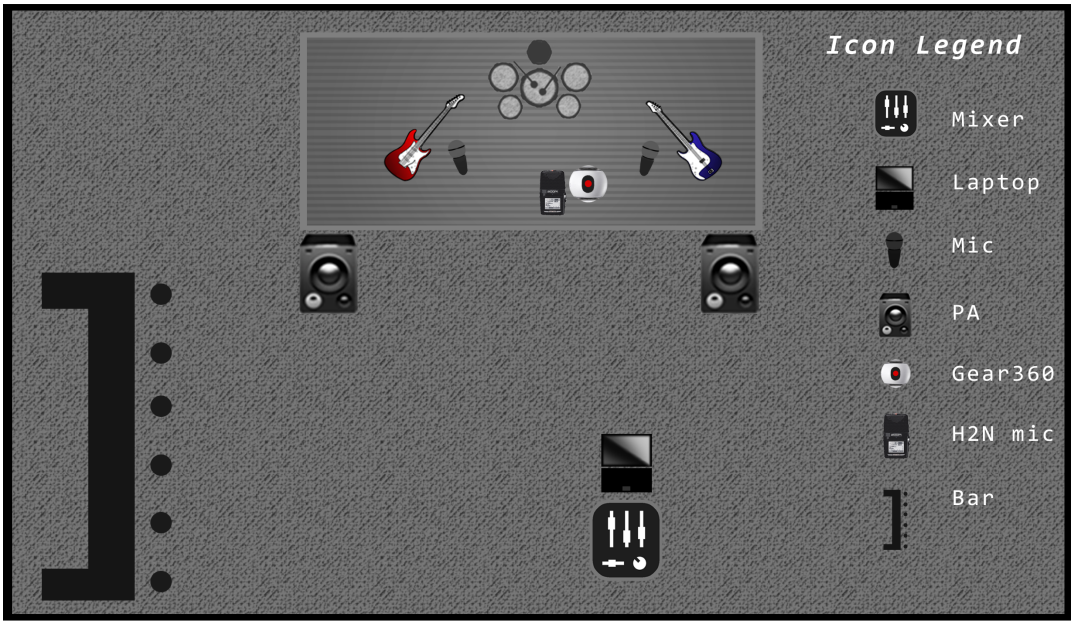

Figure 4.1: Overview of the recording setup for experiment 1

4.1.2.3 Post Production

The videos from the Samsung Gear 360 were stitched with the accompanying stitching software Action Director. The videos were further refined and processed with filters, i.e. there were lens reflections from spotlights coming from behind, and these were turned down using gamma correction in After Effects. As the Zoom H2n records directly to the B-format (ambiX ordering, explained in section 3.4.3.2), no conversion had to be made. The recordings from the mixer was processed and mixed in Reaper. The drum tracks were turned into one drum stem, and the other tracks were main vocal, backup vocal, guitar, and bass, so a total of 5 tracks. The recordings was then exported into a mono (1-channel) mix, and also exported as individual tracks which could then be placed in an FOA sound design using ATK (described in section 3.4.3.4 on ATK).

The first step was to synchronize the video with all of the audio sources. However, for this project, only a small portion of the full 2 hours recording were chosen, i.e. one song was chosen. The synchronization was done using Adobe Premiere Pro. Before importing materials into Premiere Pro, some settings had to be done so it would accept 4-channel audio files. This was done by creating a new sequence, and change the audio tracks settings to be "Multichannel" with 4 channels, and with tracktype "adaptive". The total length of the song was 06:38 minutes. The 360 degree equirectangular video was exported as an MP4 file with H.264 format, and with resolution 3840x1920, Frame Rate: 29,97, and target bitrate: 10. The audio was exported separately, so the synchronized Zoom H2N B-format track was exported as a 4 channel WAV file, the synchronized mono track was exported as a 1-channel WAV file, and the synchronized 5 individual tracks were exported as 5 individual WAV mono files. In order to make a video file that can contain FOA compatible with Gear VR, the Facebook Spatial Workstation Encoder can be used (as described in section 3.4.3.4), which requires separate sound and video files.

The synchronized 5 tracks (main vocal, backup vocal, drums, bass and guitar) had to be imported back into Reaper, such that a synthesized ambisonics mix could be created using the ATK plugin. The concept was to place the sound sources as where they appear in the visual display, i.e. the drum track are placed where the drums visually are placed, the guitar and main vocal are placed where the guitarist/singer is etc. This is of course nothing like a real concert would sound like, because the sound sources usually comes from loudspeakers that amplifies the sound in the venue. But the spatiality should be very distinguishable, and it is interesting to see how this kind of unrealistic sound representation is received at the listener, and especially in terms of presence.
In order to correctly position the individual sound sources in accordance with the visuals, another plug-in was needed that allowed the 360 video to be played and ro-

tated while placing sources. For this, the free software SpookSyncVR by SpookFM was used that applies OSC messages between Kolor GoPro VR desktop player and Reaper, such that pitch and yaw information from the Kolor GoPro VR desktop player are sent to Reaper, and they can be mapped to ATK's RotateTilt functions using automation.

When the synthesized ambisonics mix was created, it had to be exported to the ambiX format. This was also done in Reaper by routing the panned signals to a 4-channel master bus, and encoded from A format to ambiX B-format using ATK's AtoB encoder.

4.1.2.4 Implementation

Now there were three audio files ready to be muxed together with the corresponding video. The Facebook Spatial Workstation 360 software comes with an encoder that makes it possible to mux a 360 video with ambiX B-format into one MP4 file containing video and audio. In order for the spatial audio to work properly in Samsung Gear VR it had to be of the output format "Facebook 360 Video" that has the required meta data to be played on the Samsung Gear VR. This was done for the two spatial audio mixes, and the video with mono audio was rendered into a 360 video directly from Premiere Pro.

4.1.2.5 Comments on Implementation

It was a tedious procedure to ensure the synthesized mixture was correctly implemented. Even though it sounded in place when mixing in Reaper, it would sound off when played back in Gear VR. This could only be checked after mixing, exporting, muxing, transferred to a phone, then played back in VR to find out that one small thing had to be changed. However, the workflow is being improved by manufacturers as of this writing, and new possibilities arise all the time. In the mean time of this implementation being made, it was made possible to directly export 360 videos with ambiX mixes from Premiere Pro, which was not possible before. Also, the Spatial Workstation Spatializer software now contains a 360 video player, and allows an Oculus Rift or Vive to be connected such that the video can be previewed while mixing, already making the workflow much more fluent.

4.1.3 Evaluation

To evaluate how the two versions of ambisonics representations affected presence and quality of experience, a quantitative between group experiment was conducted. Furthermore, to gain insight of what kind of sound representation was the preferred listening experience for the user, qualitative interviews were conducted, where the participants listened to all three versions and ranked them after

which sound representation they preferred the most. This section will describe the methods used to gather data, followed by presentation and analysis of the data for the quantitative and qualitative data respectively. The section ends with an overall conclusion of both evaluations for experiment 1.

4.1.3.1 Quantitative Evaluation

The experiment tested on three sound representations: Two methods of FOA, one recorded with a FOA microphone, and one that was designed in an ambisonics panner plug-in using sound tracks recorded from a mixer, and lastly a mono (1-channel) recording also made with tracks from a mixer.

Conditions and Variables
The independent variable is sound representation and has 3 levels:

1. FOA using a Zoom H2n
2. FOA designed and panned using audio tracks mixer
3. Mono using audio tracks from mixer

The dependent variables are:

1. Overall Quality
2. Presence

This results in 3 test conditions with the same 360 degree video, but with a different sound representation.

Questionnaire
The questionnaire consisted of six sections:

1. General Questions
2. Motion Sickness
3. Imagery (Quality)
4. Sound (Quality)
5. Presence
6. Quality of Experience

The general questions asked for age, gender, experience with VR, and visual/audio disabilities. The motion sickness asked whether any motion sickness was experienced, and how it affected the quality of experience. The presence questionnaire was based on three of Lombard and Ditton's presence variables: Presence as Realism, Presence as Immersion and Presence as Transportation. The quality questions asked for the degree of spatiality and overall quality of both the sound and the image, and also asked for the overall quality of the experience. Both presence and quality questions were based on a Likert scale from 1-5.

Procedure
The experiment was conducted at Aalborg University in Copenhagen, A.C. Meyers Vænge 15. Participants were asked and invited to participate in the experiment. A between-group experiment method was used, meaning that different participants were used in each condition. To avoid systematic bias, the participants were randomly assigned to one of the three conditions. The conditions were labeled A (with H2n sound), B (with designed ambisonics), and C (with mono). The HMD used was a Samsung Gear VR (as presented in 3.4.6) and a pair of Philips SHB8850 headphones. The participant was asked to sit on a rotatable chair, and put on the HMD and headphones, and then asked to start the video using the touchpad on the side of the HMD. The total length of the video was 06:38 minutes, but the experience was stopped after 3 minutes. After this the participant was asked to fill out the questionnaire. An overview of the test setup can be seen in figure 4.2.

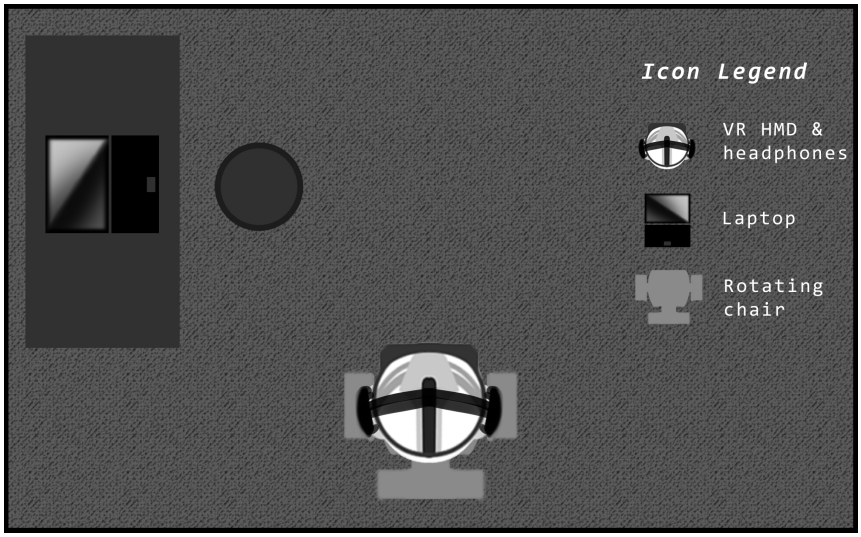

Figure 4.2: Overview of the test setup for experiment 1

Results
The results are presented in accordance with their appearance in the questionnaire. To test the differences between the three groups, the non-parametric method Kruskal-Wallis test is applied.

Statistical Test
The data has been measured on a ranked scale, i.e. a Likert scale from 1-5, and can be considered ordinal data [53]. Therefore a non-parametric statistical method is used. Non-parametric tests, in contrast to parametric tests, are often called "assumption-free", and do not follow the assumption that the data is collected from a normally distributed population [53] When dealing with a between group experiment that has three or more conditions, the non-parametric Kruskal-Wallis test can be used [54].
When performing a Kruskal-Wallis test, the first step is that all the data from each group is ranked from lowest to highest number, not caring about the group it originally comes from. In this case with number of participants N = 27 (9 in each group), it results in 27 ranked scores. The test also needs the sum of each group score. The degree of freedom (df) is k-1, where k is the number of groups, so for this test the df = 2. The significance value chosen is $\alpha = 0.05$ [54].
Based on these parameters, the test will output a chi-square number, and to reject the null-hypothesis it needs to be above a certain critical value (cv). By looking up a chi-square table, it can be find that with $\alpha = 0.05$, and df = 2, the cv = 5.99 [55]. Therefore, if the chi-square value is above 5.99, the null-hypothesis can be rejected.

1) **General Questions**
A total of 27 participants took part in the experiment (13 female, 14 male), with 9 in each condition. The average age of the participants were 24 years old. Furthermore the general questions revealed that a majority (66,7 %) of the participants had tried 360 video in VR before, and that only very few (11,1 %) consider themselves to easily experience symptoms of motion sickness. Most participants (63 %) had normal vision, and almost every participant (92,6 %) had normal hearing.

2) **Motion Sickness**
Only very few (14,8 %) experienced motion sickness during the experience, and participants rated that the quality of experience was good in regards to motion sickness with an average score of 4,18 out of 5, and a standard deviation of 0,9. This indicates that the experiences was not obstructed due to any motion sickness.

3) **Imagery**
4 questions were asked for the quality of imagery:

- *How would you rate the quality of of imagery?* [I1]
- *Did you sense spatiality in the imagery?* [I2]
- *To what degree did spatiality in the imagery contribute to the quality of your experience?* [I3]
- *Please elaborate or comment (optional)* [open question]

There were found no significant difference for any questions regarding imagery when running a Kruskal-Wallis test. The results can be seen in table 4.1. As would be expected the perceived quality of the imagery did differ significantly in any of the conditions.

Table 4.1: Table with results from the Kruskal-Wallis test on the "Imagery" questions

Question	**h**	χ^2	**p**	**mean ranks**
I1	0	2.7515	0.2527	[10.8333 14.6111 16.5556]
I2	0	0.3714	0.8305	[14.5000 13 14.5000]
I3	0	0.4318	0.8058	[12.9444 13.7778 15.2778]

4) **Sound**
4 questions were asked for the quality of sound:

- *How would you rate the quality of sound?* [S1]
- *Did you sense spatiality in the sound?* [S2]
- *To what degree did spatiality in the sound contribute to the quality of your experience?* [S3]
- *Please elaborate or comment (optional)* [open question]

There were found no significant difference for any questions when running a Kruskal-Wallis test. The results can be seen in table 4.2. Looking at the boxplot in figure 4.3 of question S1, it shows that the median is 4 in each condition, indicating that all sound representations was perceived as having a good quality.

Table 4.2: Table with results from the Kruskal-Wallis test on the "Sound" questions

Question	h	χ^2	p	mean ranks
S1	0	1.0479	0.5922	[13.8333 12.3333 15.8333]
S2	0	2.0682	0.3555	[12 16.5000 13.5000]
S3	0	3.7642	0.1523	[10 15.5556 16.4444]

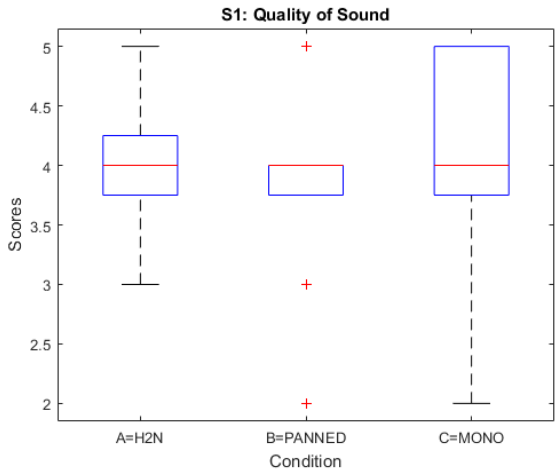

Figure 4.3: Boxplot of question S1: "How would you rate the quality of the sound?"

5) **Presence**
4 questions were asked about presence:

- *To what degree would you rate the realism of the imagery?* [P1]
- *To what degree would you rate the realism of the sound?* [P2]
- *To what degree did you feel immersed into the experience?* [P3]
- *To what degree did your feel transported into the scenery/location in the experience?* [P4]

There were found no significant difference for any of the presence questions running a Kruskal-Wallis test. The results can be seen in table 4.3.

6) **Overall Quality**
3 questions were asked about the overall quality of sound, imagery and experience:

- *Please rate the overall quality of the sound?* [QoE1]
- *Please rate the overall quality of the imagery?* [QoE2]

Table 4.3: Table with results from the Kruskal-Wallis test on the "Presence" questions

Question	h	χ^2	p	mean ranks
P1	0	2.1677	0.3383	[12 16.8889 13.1111]
P2	0	1.5234	0.4669	[11.5556 14.7222 15.7222]
P3	0	2.5369	0.2813	[12.0556 17.1667 12.7778]
P4	0	4.6764	0.0965	[9.6667 17.0556 15.2778]

- *Please rate the overall quality of the experience?* [QoE3]

There were found no significant difference for any questions regarding overall quality of sound, imagery or quality of experience running a Kruskal-Wallis test. The results can be seen in table 4.4. Looking at the boxplot of question QoE1 in figure 4.4 the synthesized ambisonics has less spread scores, whereas the mono version deviates more.

Table 4.4: Table with results from the Kruskal-Wallis test on the "Overall QoE" questions

Question	h	χ^2	p	mean ranks
P1	0	0.2194	0.8961	[13.8333 13.2778 14.8889]
P2	0	2.6937	0.2601	[11.6111 13.1667 17.2222]
P3	0	1.6528	0.4376	[11.6111 15.8333 14.5556]

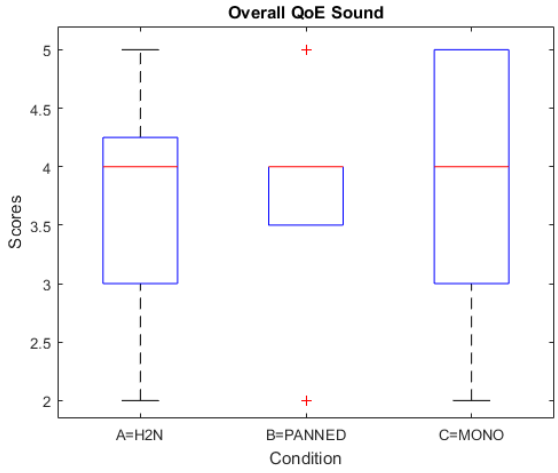

Figure 4.4: Boxplot of question QoE1: "Please rate the overall quality of the sound? QoE1"

4.1.3.2 Qualitative Evaluation

The second part of experiment 1 was a semi-structured interview. Right after the participant had answered the questionnaire in the between group experiment, s/he was asked to experience all three versions of the 360 video production, and rank the preferred versions. The participant was instructed in how to navigate in the HMD to start, stop, pause, rewind, and switch versions, and there was no limit of how many times to watch before concluding a final answer. When the participant had experienced all three versions and was ready to answer, the participant was informed that the conversation would be recorded. Two main questions were prepared:

- Which of the experiences (A,B,C) did you prefer, and why?
- Do you have other comments?

As the interviews progressed other questions appeared, mainly:

- *What did you think of the music?*
- *What did you think of having this equipment on?*

All 27 participants took part in the qualitative experiment. A transcription of all interviews can be found in Appendix B. Analysis of the data was done to find count the answers. Figure 4.5 shows a bar graph for the preferred sound experience, as well as 2nd and 3rd place. There is a strong preference towards the H2n experience, where 18 out of 27 ranked this as the most preferred. On second place the synthesized FOA mix is almost equal to the mono mix.

4.1.3.3 Findings

The quantitative results did not reveal any significant differences between the three conditions. Presence question P4 "To what degree did your feel transported into the scenery/location in the experience?" comes close with a p-value of 0.0965, where condition B with synthesized ambisonics has highest mean rank. A larger sample size might provoke better results, and also other test methods such as within group experiment, where the participant can compare different versions. This was accounted for in the qualitative data, where the results show significant preference towards the experience recorded with the FOA H2n microphone (18 out of 27).

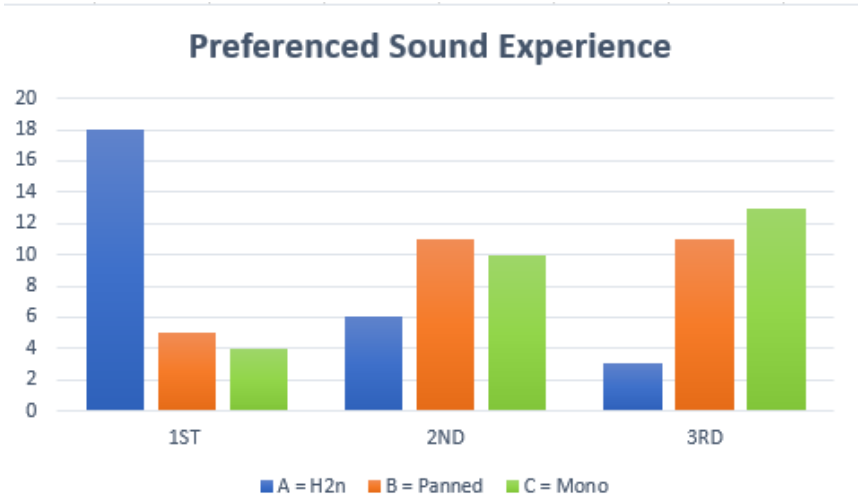

Figure 4.5: Bar graph showing preference ranks of the three sound experiences

4.2 Experiment 2

- Comparison of FOA Microphones in headphone-based VR Concerts, and the effect on Presence, Perceptual Quality and Preference

The second experiment was more refined and had more focus on the spatial sound quality attributes rather than just overall sound quality. Differing from experiment 1, the experiment used a within group design, where the participant was exposed for all versions.

4.2.1 Goal

The goal for this experiment was to evaluate the differences in perceptual spatial quality and self-evaluated presence using two FOA microphones: the full 3D AMBEO VR microphone and the low-cost horizontal only H2n microphone. The goal was also to find out which version was most preferred, and how the binaural renderings performed against mono sound.

4.2.2 VR Production

For this experiment, three different live concerts were recorded using the GoPro 7 camera rig, AMBEO VR microphone, and H2n microphone. Out of these three, only one was chosen to test on, because 3 x 3 = 9 VR experiences will quickly result in fatigueness of the test participant in a within-subject experiment, and that could possibly have an undesirable effect on the sense of presence and engagement. The reason for recording more concerts was to have more to choose from, in case of

unpredictable events, such as sudden changes in light, electrical shortage (which actually happened), or any technical failures. The chosen concert was recorded at the venue KB18 in Copenhagen with the rock band "Royal Air Force".

4.2.2.1 Preparations

For this production I had the opportunity to record with either Gear 360, GoPro 7 or GoPro 14 rig. Besides the many advantages of the Gear 360 camera, the video quality is substantially bad in low light conditions compared to the GoPro's. I chose to record with the GoPro 7 rig over the GoPro 14 rig due to the advantage of less stitching errors and less data, as described in section 3.4.5. However, there were reported problems with some of the GoPro cameras, and therefore I had to test and sort out which of the 14 that worked, and choose the 7 best.
A huge disadvantage of the GoPro rigs is that all of the cameras need to be charged in order to be sure to record more than 15 minutes. As all of the cameras need to be turned on one after another (with no light on the camera display), it should be done in good time at least 5 minutes before the concert begins, so only having 15 minutes is too unreliable to count on. The USB cables that comes with the cameras are very short, and I had to buy extension cables so they could reach down to 7 chargers on the ground. In relation to that, the AMBEO VR split cable was also too short, and I had to acquire 4 XLR cables, 6 meters long, which also enabled me to hide away from the camera rig and still monitor the audio via the Zoom F4 MultiTrack Recorder. Other purchases included strips, gaffa, batteries and a clamp to attach the Zoom H2n on a microphone stand. A full list of all the equipment and gear used can be seen on the next page.

Here is a list of all the equipment and things that were needed for the recording:

- Gaffa
- Strips
- Scissor
- Screwdriver
- Batteries
- 4 XLR cables
- Zoom F4 recorder
- AMBEO VR mic
- Zoom H2n
- 2 SD cards
- 7 micro SD cards
- Microphone stand
- 7 GoPro cameras
- 7 micro SD cards
- 360 GoPro7 Rig (Homemade)
- Fiber cloth
- Headphones (for monitoring)
- 7 USB cables + extensions
- 7 Chargers
- Zoom power adapter
- Power sockets
- Flashlight

4.2.2.2 Recording

The recording took place at the venue KB18 in Copenhagen on May 12th 2017. The band that was playing was the Danish energetic Rock/Punk band "Royal Air Force". Upon arrival, the band greeted me, and we discussed best positions of the camera. The interior had big pillars in the middle of the room, and the band was planning to jump out to the audience, so the equipment was placed in the right side fronting the scene. An overview of the setup can be seen on figure 4.6.
All cameras were set to Resolution: 1440p, Framerate: 30 FPS, Low Light: off, Sharpness: Medium, Color: GoPro, White Balance: Auto, Protune: On, and ISO 1600, because it was very dark settings. going above 1600 would result in *too* much grain. The Zoom H2n was set to its maximum recording quality when using Spatial Audio mode, 48 kHz/24 bits. The Zoom F4 (recording AMBEO inputs) was therefore also set to 48 kHz/24 bits.

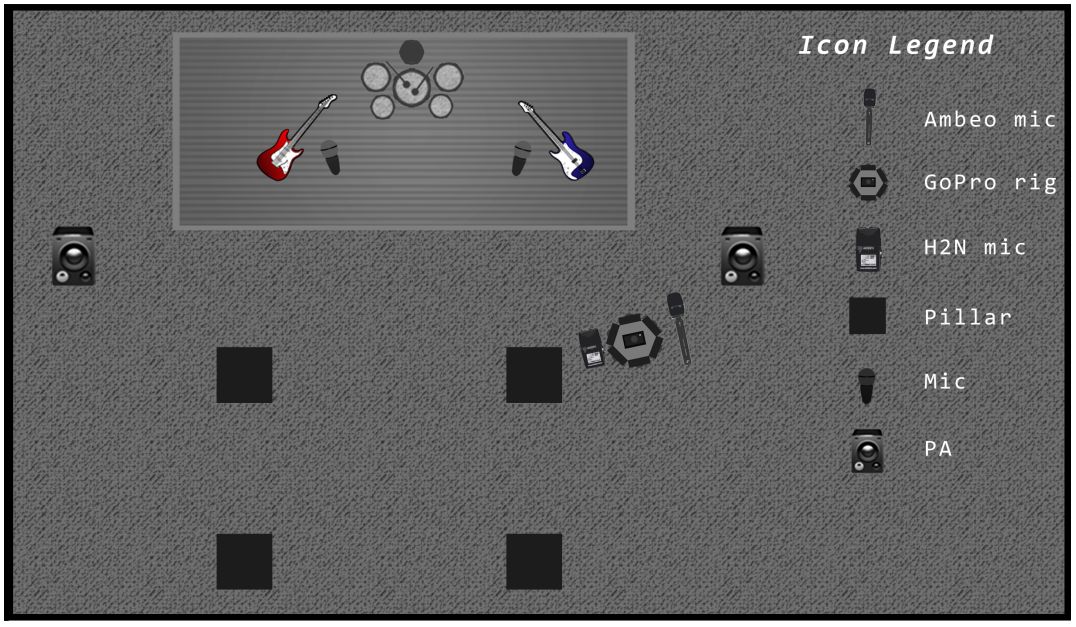

Figure 4.6: Overview of the recording setup for experiment 2

4.2.2.3 Post Production

After the recording session, all files had to be transferred to a computer. This was done by transferring data from each individual SD card, from the Zoom H2n, the Zoom F4 (which the AMBEO was connected to), and from the 7 individual GoPro cameras. As the GoPro cameras starts a new sequence for every approximately 11th minute, the videos from each camera had to be put into one long video file. This was done for every camera. The essential steps of producing the videos are

listed in steps here:

1. Transfer all files from SD cards to computer in separate folders that corresponds to the camera name/number.
2. Make a Premiere Pro file and import all videos
3. Put all videos into the timeline, so that there is 7 tracks of video and 7 tracks of audio (from the cameras) that correspond to each camera.
4. Synchronize all videos so they start at the same time
5. Export each synchronized video track into 7 individual videos
6. Import the 7 videos into the stitching software AutoPano Video Pro, and stitch the footage
7. Export the stitched videofile, and import it back into Premiere Pro
8. Synchronize stitched video with H2n and AMBEO sound tracks.
9. Export final videos

Additional post-production was done, such as removing the camera rig and cameras that was present in the camera facing downwards. This was done with the plug-in Mettle Skybox in Adobe After Effects. Some sharpening and brightness filters was also applied to the stitched video.

As the Zoom H2n converts from A to B-format internally, the sound file was ready to be imported in Premiere Pro without further do. The AMBEO VR sound, recorded with a Zoom F4 recorder, had to be converted from A to B using Sennheiser's AtoB converter (as covered in section 3.4.3.3). It was converted into the ambiX channel ordering. The only processing done to the B-format soundfiles were equal normalization and to make sure they were at equal loudness levels. A mono version was created by downmixing the AMBEO VR recording.

4.2.2.4 Implementation

The chosen HMD for this experiment was an Oculus CV1, mainly because of its ability to monitor what the participant was seeing, and simply to have more control. As mentioned, a recent update for Premiere Pro made it possible to export videos with B-format tracks, so no muxing was required. 3 videos of the concert was exported in premiere each with different sound representation: H2n B-format, AMBEO B-format, and AMBEO mono. The length of each video was 2 minutes.

4.2.3 Quantitative Evaluation

The experiment tested on three sound representations: Two methods of FOA, one recorded with a Zoom H2n, and one recorded with an AMBEO through a Zoom F4 recorder, and lastly a mono (1-channel) track made from a mixdown of the AMBEO track.

Conditions and Variables
The independent variable is sound representation and has 3 levels:

1. FOA using a Zoom H2n
2. FOA using an AMBEO
3. Mono mixdown of AMBEO

The dependent variables are:

1. Perceptual Sound Quality
2. Presence

This results in 3 test conditions with the same 360 degree video, but with a different sound representation. Preference was also asked as an open-question.

Questionnaire
The questionnaire consisted of nine sections:

1. Test Condition (filled by experimenter)
2. General Questions
3. Test 1: Sound Quality
4. Test 1: Presence
5. Test 2: Sound Quality
6. Test 2: Presence
7. Test 3: Sound Quality
8. Test 3: Presence
9. Preference

The general questions asked for gender, age, occupation, prior experience with VR, prior experience with spatial sound, motion sickness, any visual/audio disabilities, and lastly how often one attend concerts on a yearly basis.

The sound quality questions were based primarily from SAQI [25], where relevant sound quality attributes were chosen. A few additional attributes were chosen from the ITU-R BS-1284-1 inventory [26]. They were rated on a 5-point Likert scale with labels in each extreme realated to the context of the question. This resulted in nine questions. They are listed here, with the attribute category labeled in square brackets:

1. *"How would you rate the perception of where the sounds originates from?"*[Externalization]
   More internalized (inside my head) / More externalized (outside my head)
2. *"How would you rate the similarity of this sound experience with what you would expect from a live concert?"* [Naturalness (realism)]
   Low / High
3. *"How would you rate your impression of the sound sources being accurately positioned in the experience?* [Localization accuracy]
   Less accurate / More accurate
4. *"How would you rate your impression of the sound sources being placed in a way that makes the entire (sound) image balanced?"* [Directional balance]
   Less balanced / More balanced
5. *"How would your rate your impression of how clearly the sound elements (guitar, bass, singer and drums) can be distinguished from each other?"* [Clarity]
   Less clear / More clear
6. *"How would you rate the overall loudness of the sounds?"* [Loudness]
   Quit / Loud
7. *"How would you rate the intensity of reverberation (dansk: Rumklang) in the soundfield?"* [Reverberation level]
   Weak / Strong
8. *"How would you rate your impression of sound source coming in front or behind?"* [Depth-perspective]
   Not confusing / Confusing
9. *"How would you rate the intensity of perceivable artifacts (noise, clicks, disruptive sounds) in the sound experience?"* [Distortion]
   Less intense / More intense

The presence section was based on a shorter version of the ITC-SOPI questionnaire. The 44 items were shortened down to 18 questions. This was due to limit possible fatigueness the participant can get in a within-subject experiment (the participant had to answer the same questionnaire 3 times). The questions which accounted for the following 4 factors were chosen: Sense of Physical Space, Engagement, Ecological Validity, and Negative Effects.

Procedure
The experiment was conducted at Aalborg University in Copenhagen, A.C. Meyers Vænge 15. Participants were asked and invited to participate in the experiment. A within-group experiment method was used, so the same participant were used in all conditions, which were randomized with a Latin square design. The three conditions were labeled A (with H2n sound), B (with Ambeo), C and (with mono). The HMD used was a Oculus CV 1 (as presented in section 3.4.6) and a pair of Bose QuietComfort25 consumer headphones. Even though Oculus CV1 has its own pair of headphones, these were removed to use a better and high quality consumer pair instead, which also exclude surrounding obtrusive audio from the real world.
The participant was first introduced to an information sheet explaining the procedure, and then asked to fill out a consent form (see Appendix A). Then the participant were asked to fill out general questions in the questionnaire, before being exposed to the first test/VR experience (randomly assigned). After test 1, the participant was asked to fill out two sections of the questionnaire relating to sound quality and presence questions. This was repeated for condition 2 and condition 3. Lastly the participant was asked to answer which version s/he preferred. An overview of the test setup can be seen in figure 4.7.

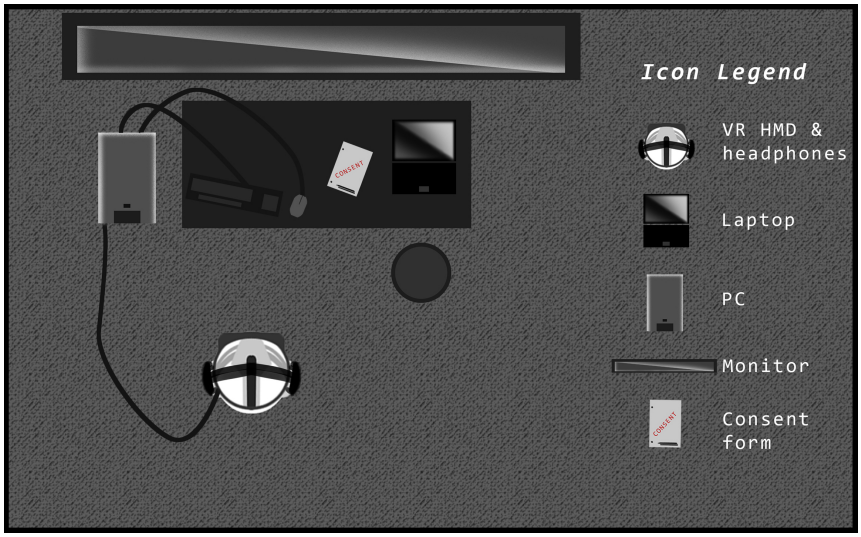

Figure 4.7: Overview of the test setup for experiment 2

Results
The results of the quantitative within group experiment will be presented here. First is a description of the statistical tests, then the questions and results are presented.

Statistical Test
The presence data that was measured followed the recommendations from ITC-SOPI, where specific questions are summed up to 4 factors of presence, by summing and calculating the means of those questions.
Dealing with non-parametric data in a within-group design, the Friedman's ANOVA test can be used for statistical tests. Friedman (non-parametric). The data is ranked from highest to lowest across conditions, and the mean ranks are outputted for each condition. If the mean ranks are fairly similar, there is probably no effect. The test statistic is denoted χ^2 (like in the Kruskal-Wallis test), and is a function of the total of the ranks in each condition, sample size, and number of conditions [54]. The significance value used will be $\alpha = 0.05$, and as the degree of freedom is 2 (conditions-1), the critical value for significant difference is 5.99 [55].

1) **General Questions**
A total of 18 participants took part in the experiment (12 male, 6 female), which means 54 in total, as the participants were reused in each condition (18 x 3). The average age of the participants was 31. 66,6 % reported they had tried VR before, and only 11,11 % had experience with spatial sound.

2) **Presence**

Friedman's ANOVA test was performed on the 4 Presence factor scores. Table 4.5 shows the result for each factor, h=1 if the null-hypothesis is rejeced, the χ^2 value, p-value and mean ranks.

Table 4.5: Table with results from the Friedman's ANOVA test on the 4 Presence factors

Presence Factor	**h**	χ^2	**p**	**mean ranks**
Sense of Physical Space	1	8.2623	0.0161	[2.5000 1.8333 1.6667]
Engagement	1	10.52	0.0052	[2.3056 2.3056 1.3889]
Ecological Validity	0	2.53	0.2821	[2.2500 1.7500 2]
Negative Effects	0	0.6047	0.7391	[1.8889 2.0278 2.0833]

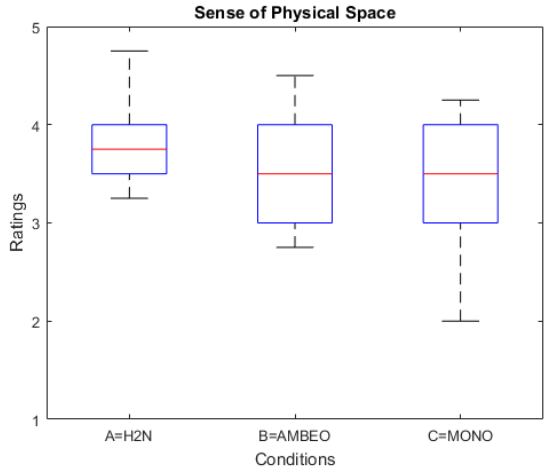

Figure 4.8: Boxplot of Presence Factor: Sense of Physical Space

The Friendman's Anova test for Sense of Physical Space rejected the null-hyopthesis at 0.05 significant level. As the Friedman test does not reveal where the difference might be, follow-up analysis is needed. By looking at the mean ranks, and also at the boxplot in figure 4.8, the largest difference appears to be between H2n and Mono. Post-hoc test is done by comparing test groups against each other with a Wilcoxon Signed-rank test. Two comparisons are made: H2n against Mono, and Ambeo against Mono. A bonferroni correction is made, so the significance value becomes $0.05/2 = 0.025$. For H2n vs Mono no significant difference were found (p = 0.0340), so the null-hypothesis is accepted. For Ambeo vs Mono no siginficant difference were found either (p = 0.5118).

By looking at the mean ranks on the Engagement factor, the difference might lie between either H2n (2.3056) and Mono (1.3889) or Ambeo(2.3056) and Mono. The

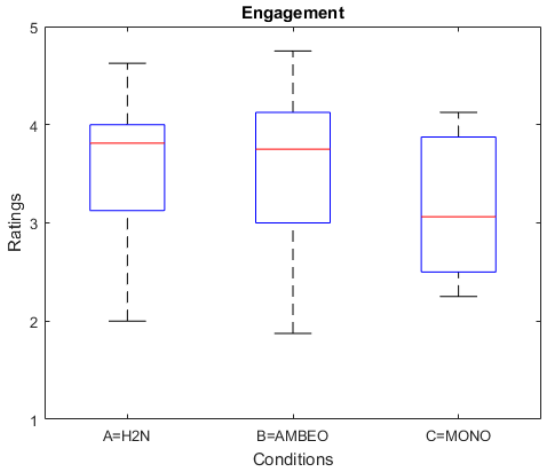

Figure 4.9: Boxplot of Presence Factor: Engagement

boxplot in figure 4.9 Bonferroni correction was again applied to 0.025. Results from the Wilcoxon tests show no significant difference for H2n vs Mono (p = 0.0753), and no significant difference for Ambeo vs Mono (p = 0.0613).

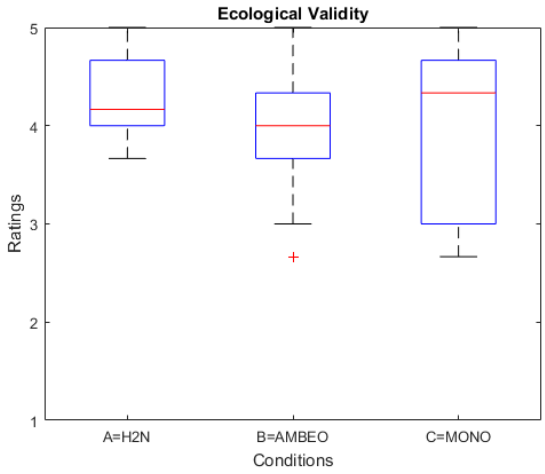

Figure 4.10: Boxplot of Presence Factor: Ecological Validity

The results for Ecological Validity showed no significant difference. Looking at the boxplot in figure 4.10, the mean scores in all conditions were generally very high, although there are more variation in the Mono condition. The results for Negative Effects did not show any significant difference, and generally mean scores in all conditions were low, which indicates that no bad influences, e.g. motion sickness, was experienced.

3) **Sound Quality**

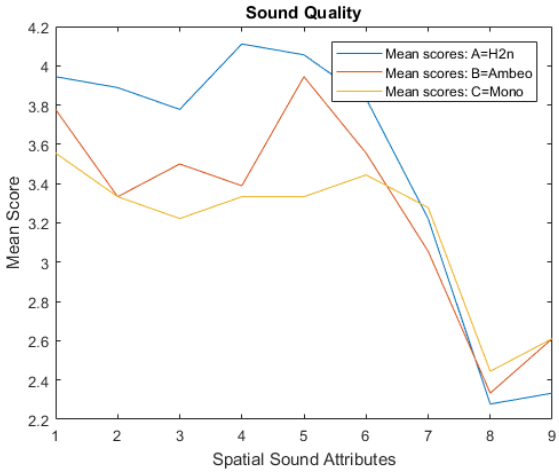

Figure 4.11: Mean score plot of the 9 sound quality questions

Figure 4.4 shows a plot of mean scores for the 9 sound quality questions for each condition. In general most ratings for all conditions lie between 3.2 and 4.1, expect from the last two questions relating to front-back confusion and distortion (with low ratings being good). Interestingly the H2n condition have generally higher ratings than both Ambeo B-format and Ambeo Mono. The largest difference is seen for question 4 regarding balance in the sound image. In question 2 about naturalness or realism, ambeo and mono have equal scores.

4) **Preference**
The last question in the questionnaire asked the participant which version s/he preferred. Figure 4.12 shows a diagram of the preferences.

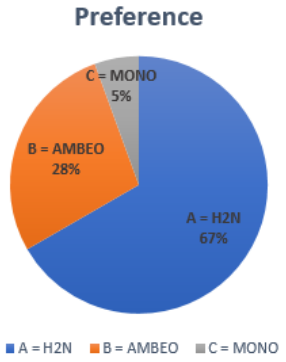

Figure 4.12: Diagram showing preference of the three versions presented in percantages

Like in experiment 1, there is surprisingly a preference towards the H2n FOA representation.

4.2.3.1 Findings

Using Friedman's Anova test there were found significant differences in presence factors Sense of Physical Space and Engagement. However through post-hoc analysis, the null-hypothesis could not be rejected for any of them, although H2n vs Mono was close to being rejected in Sense of Physical Space with p = 0.0340, at a significance value of 0.025.
The spatial audio quality questions showed that the H2n generally had slightly higher scores than the others, but not at any significant level. Most interestingly H2n was rated as having quite more balance, better position and loudness. Another interesting observation is that for the question regarding externalization, the mono recording is perceived as coming more "out-of-the-head". Also in question 7 about reverberation, the scores in all conditions are almost neutral (scoring of 3), which could indicate that the participants did not understand the question. The same goes for question 8 about front-back confusion. All in all there were not rated any disturbing factors or artifacts for any of the conditions.

Chapter 5

Discussion

The two experiments conducted did not show any significant differences in regards to presence nor quality. In the qualitative evaluation of experiment 1 (N=27), participants had a strong preference towards H2n, where 18 rated this as 1st place out of the 3 versions. However, participants reported verbally that they felt it was a bit louder, and more balanced (in regards to levels of instruments) than the others. Others reported that there were some artifacts in the synthesized version. It can be concluded that the H2n did not ruin the quality of experience, and that FOA can successfully be implemented in VR concerts, without concern of consumers accustomed to industry stereo/mono standard.
For the two presence factors in experiment 2 "Sense of Physical Space" and "Engagement" a significant difference was almost found ($p = 0.034$, and ($p = 0.075$) between H2n and Mono conditions. With a sample-size of N=18 for each condition, larger samples should be assessed to show a more valid indication of significance. Experiment 2 (N=18) compared the low cost H2n microphone with the AMBEO VR mic in terms of perceptual quality of naïve listeners. The results surprisingly showed slightly higher rates towards H2n. This does not equal that H2n is superior in any way. It appears that the naïve listeners might have not clearly understood all questions that well, and the small sample size also is not indicative of any conclusions. However, when asking the participants' about their preferred version, 67 % says H2n. In accordance to the findings in [11], Ambeo and H2n are not so distinguishable to listener's, and as such despite its low price, the H2n seem to work sufficiently for FOA implementation in VR concerts.

Chapter 6

Conclusion

In this study several implementation of first-order ambisonics in VR concerts have been explored, and how they affect presence and subjective quality. Findings did not show any significant differences between ambisonics and mono renderings, neither between different ambisonics renderings. However preference ratings show for the two conducted experiments that participants preferred the ambisonics renderings recorded with a low-cost Zoom H2n microphone.

Presence ratings were generally high in all conditions through both experiment, showing that participants did feel high degree of presence, but not that ambisonics significantly increased presence. Perceptual quality attributes showed slightly higher scores for H2n as opposed to Mono and AMBEO, but not significantly higher.

Bibliography

[1] Heather Bellini. "Understanding the race for next computing platform". In: *Profiles in Innovation: Virtual & Augmented Reality* (Jan. 2016), p. 8. URL: http://www.goldmansachs.com/our-thinking/pages/virtual-and-augmented (Date accessed: 05/02/2017).

[2] Abesh Thakur. *Introducing Spatial Audio for 360 Videos on Facebook [ONLINE]*. Oct. 2016. URL: https://media.fb.com/2016/10/07/introducing-spatial-audio-for-360-videos-on-facebook/ (Date accessed: 05/02/2017).

[3] Neal Mohan. *One step closer to reality: introducing 360-degree live streaming and spatial audio on YouTube[ONLINE]*. Apr. 2016. URL: https://youtube.googleblog.com/2016/04/one-step-closer-to-reality-introducing.html (Date accessed: 05/02/2017).

[4] Claudia Hendrix and Woodrow Barfield. "The sense of presence within auditory virtual environments". In: *Presence: Teleoperators and Virtual Environments* 5.3 (June 1996), pp. 290–301.

[5] Daniel Vastfjall Ana Tajadura-Jiménez & Mendel Kleiner Pontus Larsson Aleksander Valjamae. "Auditory-Induced Presence in Mixed Reality Environments and Related Technology". In: *The Engineering of Mixed Reality Systems*. London: Springer London, 2010, pp. 143–163.

[6] Daniel Vastfjall. "The subjective sense of presence, emotion recognition, and experienced emotions in auditory virtual environments". In: *Cyberpsychology & behavior: the impact of the Internet, multimedia and virtual reality on behavior and society* (Apr. 2003), pp. 181–188.

[7] M. Kobayashi, K. Ueno, and S. Ise. "The Effects of Spatialized Sounds on the Sense of Presence in Auditory Virtual Environments: A Psychological and Physiological Study". In: *Presence* 24.2 (May 2015), pp. 163–174.

[8] G. Catusseau C. Guastavino V. Larcher and P. Boussard. "Spatial audio quality evaluation: comparing transaural, ambisonics and stereo". In: *Proceedings of the 13th International Conference on Auditory Display*. Montreal, Canada, June 2007.

[9] Fábio Wanderley Janhan Sousa. "Subjective Comparison between Stereo and Binaural Processing from B-Format Ambisonic Raw Audio Material". In: *Audio Engineering Society Convention 130*. May 2011.

[10] Chris Baume and Anthony Churnside. "Upping the Auntie: A Broadcaster's Take on Ambisonics". In: *Audio Engineering Society Convention 128*. May 2010.

[11] Enda Bates et al. "Comparing Ambisonic Microphones—Part 2". In: *Audio Engineering Society Convention 142*. May 2017.

[12] Mel Slater. "A note on presence terminology". In: *Presence connect*. Vol. 3. Jan. 2003.

[13] Matthew Lombard et al. *Immersed in Media: Telepresence Theory, Measurement & Technology*. Springer Publishing Company, Incorporated, 2015.

[14] Marvin Minsky. "Omni". In: *Presence: Teleoper. Virtual Environ.* 12.5 (Oct. 2003), pp. 456–480.

[15] Matthew Lombard and Theresa Ditton. "At the Heart of It All: The Concept of Presence". In: *Journal of Computer-Mediated Communication* 3.2 (1997).

[16] Bob G. Witmer and Michael J. Singer. "Measuring Presence in Virtual Environments: A Presence Questionnaire". In: *Presence: Teleoper. Virtual Environ.* 7.3 (June 1998), pp. 225–240.

[17] Frank Biocca, Chad Harms, and Judee K. Burgoon. "Toward a More Robust Theory and Measure of Social Presence: Review and Suggested Criteria". In: *Presence: Teleoper. Virtual Environ.* 12.5 (Oct. 2003), pp. 456–480.

[18] T. Furness III Maxwell J. Wells J.D. Prothero D.E. Parker. "Towards a robust, quantitative measure for presence". In: *Experimental Analysis and Measurement of Situation Awareness*. Washington, 1995, pp. 359–366.

[19] Daliza Crane Bill Davis Gisela Gil-Egui Karl Horvath & Jessica Rossman Matthew Lombard Theresa Ditton. "Measuring presence: A literature-based approach to the development of a standardized paper-and-pencil instrument". In: (2000).

[20] Martin Usoh et al. "Using Presence Questionnaires in Reality." In: *Presence: Teleoperators & Virtual Environments* 9.5 (2000), pp. 497 –503.

[21] Jane Lessiter et al. "A Cross-Media Presence Questionnaire: The ITC-Sense of Presence Inventory". In: *Presence: Teleoper. Virtual Environ.* 10.3 (June 2001), pp. 282–297.

[22] S. Bech and N. Zacharov. *Perceptual Audio Evaluation - Theory, Method and Application*. Wiley, 2007.

[23] Francis Rumsey. "Spatial Quality Evaluation for Reproduced Sound: Terminology, Meaning, and a Scene-Based Paradigm". In: *J. Audio Eng. Soc* 50.9 (2002), pp. 651–666.

[24] Nick Zacharov and Kalle Koivuniemi. “Unravelling the Perception of Spatial Sound Reproduction: Analysis & External Preference Mapping”. In: *Audio Engineering Society Convention 111*. Nov. 2001.

[25] Alexander Lindau. *Spatial Audio Quality Inventory (SAQI). Test Manual.* Tech. rep. Technische Universität Berlin, Feb. 2014. URL: http://dx.doi.org/10.14279/depositonce-1.2 (Date accessed: 05/10/2017).

[26] ITU-R BS.1284. *General method for the subjective assessment of sound quality*. International Telecommuncation Union - Radiocommunication Sector. 1997. URL: https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.1284-1-200312-I!!PDF-E.pdf (Date accessed: 05/10/2017).

[27] Johannes M. Zanker. *Sensation, perception and action: An evolutionary perspective*. Palgrave Macmillan, 2010.

[28] J. Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization*. MIT Press, 1997.

[29] Francis. Rumsey. *Spatial Audio*. Music technology series. Focal Press, 2001.

[30] E. Colin Cherry. “Some Experiments on the Recognition of Speech, with One and with Two Ears”. In: *The Journal of the Acoustical Society of America* 25.5 (1953), pp. 975–979.

[31] H. Mcgurck and J. W. Macdonald. “Hearing lips and seeing voices”. In: *Nature* 264.246-248 (1976).

[32] W. R. Thurlow and C. E. Jack. “Certain determinants of the "ventriloquism effect"”. In: *Perceptual & Motor Skills* 36 (1973), pp. 1171–1184.

[33] Steve Savage. *The Art of Digital Audio Recording*. Oxford University Press, 2011.

[34] C. I. Cheng and G. H. Wakefield. “Introduction to head-related transfer functions (HRTFs): Representations of HRTFs in time, frequency, and space”. In: *AES: Journal of the Audio Engineering Society* 49.4 (Apr. 2001), pp. 231–249.

[35] Michael A. Gerzon. “Periphony: With-Height Sound Reproduction”. In: *J. Audio Eng. Soc* 21.1 (1973), pp. 2–10.

[36] Matthias Frank, Franz Zotter, and Alois Sontacchi. “Producing 3D Audio in Ambisonics”. In: *Audio Engineering Society Conference: 57th International Conference: The Future of Audio Entertainment Technology – Cinema, Television and the Internet*. Mar. 2015. URL: http://www.aes.org/e-lib/browse.cfm?elib=17605.

[37] Dave G Malham. “Spatial hearing mechanisms and sound reproduction”. In: *University of York* (1998).

[38] Christian Nachbar; Franz Zotter; Etienne Deleflie; Alois Sontacchi. "AmbiX – A Suggested Ambisonics Format". In: *Ambisonics Symposium 2011*. Lexington (KY), June 2011. URL: http://www.ambisonictoolkit.net/assets/files/2014-ICMC-ATK-Reaper.pdf.

[39] Lucas Matney. *Facebook just bought VR audio company Two Big Ears and is making their tech free to developers [Online]*. May 2016. URL: https://techcrunch.com/2016/05/23/facebook-just-bought-vr-audio-company-two-big-ears-and-is-making-their-tech-free-to-developers/ (Date accessed: 05/18/2017).

[40] Angelo Farina. *Conversion form Ambisonics to TBE format [Online]*. 2017. URL: http://pcfarina.eng.unipr.it/TBE-conversion.htm (Date accessed: 05/18/2017).

[41] F. Rumsey and T. McCormick. *Sound and Recording*. Elsevier/Focal, 2009.

[42] Sennheiser. *AMBEO VR MIC. [Online]*. 2016. URL: https://en-us.sennheiser.com/microphone-3d-audio-ambeo-vr-mic (Date accessed: 05/20/2017).

[43] Zoom. *HOLLYWOOD SOUND GOES INDIE: THE ZOOM F4. [Online]*. 2017. URL: https://www.zoom.co.jp/products/field-recording/f4-multitrack-field-recorder (Date accessed: 05/20/2017).

[44] Zoom. *THE WORKHORSE OF FIELD RECORDERS: THE ZOOM H2N. [Online]*. 2017. URL: https://www.zoom-na.com/products/field-video-recording/field-recording/zoom-h2n-handy-recorder (Date accessed: 05/20/2017).

[45] Trond Lossius and Joseph Anderson. "ATK Reaper: The Ambisonic Toolkit as JSFX plugins." In: *ICMC*. Michigan Publishing, 2014. URL: http://www.ambisonictoolkit.net/assets/files/2014-ICMC-ATK-Reaper.pdf.

[46] Adam Mckeag and David Mcgrath. "Sound Field Format to Binaural Decoder with Head Tracking". In: *Preprint of the Audio Engineering Society for the 6th Australian Regional Convention* 4302 (1996).

[47] G.C. Burdea and P. Coiffet. *Virtual Reality Technology*. Academic Search Complete vb. 1. Wiley, 2003.

[48] Jonathan Steuer. "Defining Virtual Reality: Dimensions Determining Telepresence". In: *Journal of Communication* 42 (1992), pp. 73–93.

[49] W.R. Sherman and A.B. Craig. *Understanding Virtual Reality: Interface, Application, and Design*. Morgan Kaufmann series in computer graphics and geometric modeling. Morgan Kaufmann, 2003.

[50] Paul James. *HTC Vive Review: A Mesmerising VR Experience, if You Have the Space [ONLINE]*. Apr. 2016. URL: http://www.roadtovr.com/htc-vive-review-room-scale-vr-mesmerising-vr-especially-if-you-have-the-space-steamvr/ (Date accessed: 04/21/2017).

[51] Oculus VR. *Rift+Touch [ONLINE]*. 2017. URL: https://www.oculus.com/rift/ (Date accessed: 04/21/2017).

[52] Samsung Electronics. *Gear VR [ONLINE]*. 2017. URL: http://www.samsung.com/global/galaxy/gear-vr/ (Date accessed: 04/21/2017).

[53] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. *Research Methods in Human-Computer Interaction*. Wiley Publishing, 2010.

[54] A.P. Field and G. Hole. *How to Design and Report Experiments*. SAGE Publications, 2003.

[55] Bryan E. Denham. "Chi-Square Table". In: *Categorical Statistics for Communication Research*. John Wiley & Sons, Ltd, 2016, pp. 259–260. URL: http://dx.doi.org/10.1002/9781119407201.app1.

Appendix A

Consent Form

See next page.

Participant consent form

Title of Project: Exploration of First-Order Ambisonics Usage in VR Concert Experiences

Name of Researcher: Simone Patricia Vinkel

Please tick the boxes

1. I confirm that I have read and understand the information sheet for the above study. I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily. ☐

2. I understand that my participation is voluntary and that I am free to withdraw without giving any reason. ☐

3. I understand that direct quotes may be used anonymously from the questionnaire in any write up, presentation or publication of this study. ☐

4. I understand that all personal information will be anonymised and treated confidentially. ☐

5. I agree to take part in the above study. ☐

Name of Participant …….

Signature …….. Date …….

Name of Researcher ……

Signature …….. Date ……

Appendix B

Transcription - Experiment 1

See next page.

I = INTERVIEWER
P# = PARTICIPANT

Participant 1 (first experience A):

I: Which of the experiences (A,B,C) did you prefer, and why?

P1: B. A wasn't loud enough. Not loud enough, but the sound quality was not quite in depth. Intensity of sound was low, and C was too distorted. So B was like inbetween and the best.

I: Do you have other critical comments?
P1: No any other comments

I: Could you eventually also rate the second and third? [...] You had one that was clearly your favorite, one was that?
P1: B
I: can you also elaborate what is second best and third best?
P1: Yea I would prefer B, A and C. Because C, the sound was distorted.

Participant 2 (first experience B):

I: Which of the experiences (A,B,C) did you prefer, and why?
P2: I would prefer A - because it was more like equivalent, if you want to hear a song. The quality was more likely to sound like that, unlike C - which was like if you was a part of the experience. And B, there was something when I turned around, there happened something with the sound. It was like I was looking away, and then. It was like a mixture of C and A.. B, I think. If it was the sound I would prefer A. it was the experience I would probably prefer B.
I: With the sound in mind, you say you prefer A? If A is the first one, then what's the second place and third place?
P2: Then it would be B and then C.

I: Do you have other critical comments?
P2: I have never tried VR before, so it was quite funny to try. Actually I think it was quite fun with the sound, when you looked away, it felt like it came from behind – that is the sound was toned down.

Participant 3 (first experience C):

I: Which of the experiences (A,B,C) did you prefer, and why?
P3: I think I preferred A. And I also think I answered wrong, actually (the questionnaire), because when I saw, i think it was C i saw first.
I: Yes
P3: And then I fought I could hear like okay so she sang from over here, and the bass came from over here. But now when I could like compare them, I could hear that in the A, so here is the voice, here is the bass, but when I tried to listen to number C, it sounded like it was more like all of it were together. I think maybe that was why I preferred A. But number C was also

good, but A was more like a real concert thing, maybe, because you could actually hear oh, what's happening over here now, she sing, oh here came the bass.
I: Okay, so you preferred A?
P3: Yes. And maybe i'm totally wrong, but it sounded like that!
I: If A is the best, which comes second?
P3: I think it would be C then.

I: Do you have other critical comments?
P3: No - it was fun and cool! I liked the music.

Participant 4 (First experience C):

I: Which of the experiences (A,B,C) did you prefer, and why?
P4: I think I preferred the B. Not quite sure why, but I think I thought the sound just came in better for me, and a more space way that i thought I could hear it from more sides.
I: If B was the one you preferred the most, then was was your second choice?
P4: Probably C, and then A.
I: And why?
P4: Oh yeah, that's always a good question. I preferred C, and why even though? I think I just thought I could hear the sound better in that one. Then on A, where I just didn't feel quite as immersed as I did in the other two. I definitely felt most immersed in B, and then C. I didn't really feel that immersed about it in A.

I: Do you have other critical comments?
P4: I would say the picture quality is okay, but not completely good for me. But I've also got bad eyes, so that probably don't help on that a lot.
So for me I couldn't get the spatial awareness, which made me really go to the sound for immersion instead. And I think B did that most for me.

Participant 5 (First experience B):

I: Which of the experiences (A,B,C) did you prefer, and why?
P5: A! I think the sound was more clear to me, compared to the other two, where the last one was a bit more about the instruments. It was too noisy.
I: The last one, which one was that?
P5: C.
I: If A was the best, then what is the second one for you?
P5: B. A, B, C.
I: and why is b the second?

P5: Because I think it was a mix of the two in my opinion.

I: Do you have other critical comments?
P5: It was great!

Participant 6 (First experience A):

I: Which of the experiences (A,B,C) did you prefer, and why?
P6: That would be A. A is number one. It felt that the spatiality of the sound was the strongest out of all of them. It kind of made it more real when I was turning my head, cause i could really hear things coming from sides, when I have my head sideways.
I: If you should choose one of the 3 sound regimes, you would choose no. A? Due to its quality?
P6: Yea, I would choose A. Yea, quality and spatial. The 3D sound was very nice.
I: Which one was your second favorite?
That would be B. It felt odd though, compared to the first one. I could still hear 3D effect, I don't think it was as strong. Mostly the reason why I rank it second, because it kind of leads me to the worst part which was C, i think it was just stereo, and it didn't change at all. And it it felt really weird when i started turning my head, so that's why i didn't like it so much.

I: Do you have other critical comments?
P6: nah

Participant 7 (first experience A):

I: Which of the experiences (A,B,C) did you prefer, and why?
P7: [Long pause] That's kind of hard, because.
I: Based on the sound experience
P7: I guess the second one then. I had a little bit of a trouble to actually hear the difference between the first one and the second one. So it's kinda hard for me to pick soundwise, but the third one was like.. It didn't sound like it had any panning. So I guess that kinda took you a little bit out of the immersion because the sound was exactly the same. But the first one, two ones sounded the same to me. So I wouldn't know which one to pick.
I: Which one would you take?
P7: Two I guess, so B.
I: Which was the second then?
P7: For sound? I guess, C. Because, I actually like to get the sound like it should be. When it's not like in the immersive way I guess.1
B. second preference, C - sound like it should be not like in the immersive way.
I: And then the third spatial sound experience as the last?
P7: Yeah.

I: Do you have other critical comments?
P7: Not beside that weird line that was in the middle of it.. I also noticed that it was recorded differently, the version C. Like it felt like everything was closer, and the others was further away.

I = INTERVIEWER
P# = PARTICIPANT

But I'm not really sure which one I preferred. Because it was easier to see what was going in the on where i was a little closer, right, but not sure that it was more realistic. Because not really sure how big the roomscale was. It was a bit weird to be on top of the scene. Probably makes more sense to me front audience or something. Besides that i'm not really a live concert guy.
I: did you enjoy the music?
P7: Not really my genre.

Participant 8 (First experience C):

I: Which of the experiences (A,B,C) did you prefer, and why?
P8: I think the B. Because the C, the sound is very bad.
I: Bad how?
P8: Very inside the head, very internalized. There is no spatialization. And then for the.. I don't know. For the A, there was like jumping of the sound, when I was moving. And for the B, that was better. But I found like a spot where it was like if it was a low pas filter. It was weird, I don't know. If i looked between in the scenario close to the guitarist. It was weird. But anyway, I think the B.
I: If B is the best, what is the second best?
P8: A.
I: What did you think of the music?
P8: I liked it.

I: Do you have other critical comments?
P8: After watching the first video, and then I listened the other ones and thought "Yea okay, that's much better…!"

Participant 9:
I: Which of the experiences (A,B,C) did you prefer, and why?
P9: B. Because the others sound bad. Cause the headtracking thing. Can't remember what it's called anymore. Where I looked the sound came from the right place.
I: So be was the best. What was the second best?
P9: A. Because C sounded crap.
I: Why did C sounded like crap?
P9: It was really tinny

I: What did you think of the music?
P9: It's allright, good.

I: Do you have other critical comments?
P9: No, but B was miles better.
I: How much better? Or in what terms what is better?
P9: It actually fit, where the other ones was static, I guess.

I = INTERVIEWER
P# = PARTICIPANT

Participant 10:

I: Which of the experiences (A,B,C) did you prefer, and why?
P10: I prefer C. It sounds more real, A sounds a bit hollow, B sounds like you are actually wearing headphones. If you can put it like that.
I: What's the second place?
P10: Second place. B.
I: Why is that better than A?
P10: A sounds like you are out in space, and I generally i just don't like that very distant sound. I don't know how to express in any other ways. But it sounds very distanced.
I: What did you think of the music?
P10: That was nice.

I: Do you have other critical comments?
P10: [Long pause.] One thing is the glasses. They are quite heavy. Another thing, it would be nice if you could actually - i don't know if that's something you can do anything about, but if you could move around in the room, instead of being that close to the stage. And then also, the sound, it's a bit confusing when I can see she's standing to the left, but the sounds sounds like it comes straight in front of me. So i would maybe expect it to come from the left.

Participant 11:
I: Which of the experiences (A,B,C) did you prefer, and why?
P11: I guess A, because it gives me the choice. You can hear it stronger, higher volume. That's what I felt, maybe i'm just confused. Then i thought C, there's no stereo, no spatiality in the sound, and then B, there is spatial, but you don't choose who you want to hear more. I think. Again, that's what I heard.
So, A is my favorite. B in the middle. C the least favorite.
I: What do you think of the music?
P11: I like the music, it's nice music

I: Do you have other critical comments?
P11: Yea, I think it's not very good because of the whole thing on the head. If you can deal with that, if you have something smaller that does that, then it's good. Another thing is that the whole concert experience, yea you have to be there with the rest of the crowd in order to actually get the full experience and actually make it enjoyable in my opinion.
I: When you say have it smaller, what do you mean?
P11: Like the size of my glasses, you know, really small. It needs to really jump in technology.
I: did you feel the headphones were in the way?
P11: I didn't really feel the headphones. They are usually really light, but this thing, the VR box, is kinda heavy. Can't really enjoy it, move around with it, dance to it with, you know. You just have to sit or stand for this. And, you know, at the end you're just like oh, i just want this to get overwith, with all these things on my head.

I = INTERVIEWER
P# = PARTICIPANT

Participant 12:

I: Which of the experiences (A,B,C) did you prefer, and why?

P12: The third one, C. When I scrolled through all of them, the first one (A), felt like it was okay, I could get immersed into the whole scenary. The second one felt more like a recording, like a studio recording. I could hear the audio was tempered with. The third one, it felt like I was actually at the concert. Like with the echo, and everything, I felt that there was more of an echo, and better sound on the instruments rather than the vocalist. So the third one got me into the scenary a lot more. Based on prior experience, because I've been to concerts in open field and in closed. So I could relate more to the third one.

I: If the third one was the best, what was the second best?

P12: A. From a concert point of view, it was A because it still got me immersed, like I could hear the craft of the instrument, yea. Like if you compare it to C, instruments was up here, vocalist here. A, it was like middle ground, and B just felt like a studio recording. If you get what I mean.

I: What did you think of the music then?

P12: Like the genre? It was relaxing, really appropriate for a bar. Yea, mostly closed space. I could definitely drink a beer there, that's for sure.

I: Do you have other critical comments?

P12: The whole experience? The sound was good, I could relate to it. The imagery was a bit poor, and the resolution wasn't that great. Considering my field of view, i could only have a focus point right in the middle, so i had to turn my head to look at each band member. That's the thing that lowered my experience. The resolution.

I: What about the size of the gear?

P12: I felt it was a bit heavy on the front. But that's nothing new.

I: What about the headphones?

P12: They were pretty comfortable.

Participant 13:

I: Which of the experiences (A,B,C) did you prefer, and why?

P13: I think i preferred the C. It felt a little more realistic somehow. They sounded very similar, but the C was more like being there. It's hard to explain. Maybe because of the rumklang.. The reverb.

I: Okay. So if C is the one you prefer, what is the second place?

P13: I think that is A maybe. So B is the last.

I: Why is A better than B?

P13: Hmm. The sound was a little bit more clear, I think.

I: What did you think of the music?

P13: Like the style, the song? It was okay I think. Not something I would listen to myself, but I liked the song.

I: What about having this gear on?

P13: That was nice, fine. Nothing about that.

I = INTERVIEWER
P# = PARTICIPANT

I: Do you have other critical comments?
P13: No, I think it was fun to try. I had no motion sickness or anything, so that was good.

Participant 14:
I: Which of the experiences (A,B,C) did you prefer, and why?
P14: A, because there was much more stereo. I could hear the guitar and woman singing here, and bass and singing hear - it felt much more spacious.
I: What was your second best then?
P14: The C. Because there was no like, when I looked around it moved around with it, but I just felt like you could hear everything as a whole like very good.
I: And you didn't feel that about B?
P14: No, the B, I'm not quite sure. It was a bit weird the B one, the sound.
I: How weird?
P14: I don't quite remember, but when I looked around, the sound was not where I wanted it to be.
I: What did you think of the music then?
P14: The music? That was nice.
I: What kind of music do you listen to normally?
P14: All sorts of music - no like specific. But I listen to a lot of rock and music like that actually.

I: What did you think of having this equipment on?
P14: I think it's nice - I think it's always fun to try VR.

I: Do you have other critical comments?
P14: The video quality was maybe a bit low. That's the main thing, but that might be the headset. Because I feel like I could see the pixels at all times. So that was a bit confusing and a bit annoying. But it's still quite good, but I think that's the only thing I really had.

Participant 15:
I: Which of the experiences (A,B,C) did you prefer, and why?
P15: [Uhm..] I liked, I think video C, I think. The one where, if you turn around and looked at the bass player, you could only mostly hear the bass, and a little bit guitar, and then you looked around to the guitar, and you could only hear mostly the guitar player, and not the bass player. That's the video C right? No, was it video B?
I: yea, I think it's B.
P15: That was pretty cool. But audio wise, I liked A actually, better. Quality-wise. But I like the idea that if you're looking at the bass player there's mostly bass, and guitar players, mostly guitars. But I think it wasn't that realistic, because they were standing so close to each other, so they shouldn't cancel out, they should be same volume, when they are standing that close. I think it would work if they were standing like 100 meters to each side, and you walk towards them. But yeah. The sound in A was preferable.
I: So A was the best, and then..?

P15: C, and then B. B was the one with the different audio? Right? I think so. [...] So the one where you looked at the bass-player were the loudest, and that's the least I think.
I: What did you think of the music then?
P15: The music? I think it's all right. Standard rock to be.
I: What kind of music do you listen to normally?
P15: Hip-hop mostly, and RnB.
I: What do you think of having the equipment on?
P15: It's not that bad. It doesn't really interfere with the experience, if that's what you're asking. It's not that heavy. I got these big headphones at home, so it's not that bad, I'm kinda used to it. And the glasses is not that heavy compared to any other models there is, you know HTC vive and oculus are kinda heavy, this is lighter, which makes it easier to actually use it for a longer time. So it's not that bad.

I: Do you have other critical comments?
P15: I also wrote it in the questionnaire, but I would like to at least see some of the audience - you could only see those at the bar. So I felt kinda alone, it would be pretty cool to rock out with the audience and see their reaction. Because they are also the whole experience in a concert is the audience as well. And here you're just alone, and that's a big problem I think. Kinda feel like this more intimid, and that rock-on. So. Yea I think that's the only problem so far.

Participant 16:
P16: [...]
I: Which of the experiences (A,B,C) did you prefer, and why?
P16: [...] I think A was the better one, I just think the sound was better, what do you call.. klang?
I: tone?
P16: What about the second best?
I: Hmm, I think it was number 3, but that might be because I had so many problems with number B. But can't tell if that has affected it. But my overall experience was the number 3 was better than number 2, and that A was the best.
I: So…
P16: A,C,B.

What do you think of the music that you heard?
It was nice. I like that kind of music, cozy.
Is it something that you listen to normally?
I don't listen to that kind of music a lot. I listen to a little more hard rock. But I generally like a lot of different kind of music, so this kind of music is nice for the right situation. Not too loud, it's cozy and easy to listen to when you are with other people for example.

What do you think of the experience, like, what do you think of having the equipment on?
That's fine for me. I like the ability to distract from everything, kinda just fall into this experience. Don't get distracted as easily. You a kind of immersed in it.

I = INTERVIEWER
P# = PARTICIPANT

I: Do you have other critical comments?
P16: I think the image quality could have been better. Very pixilated. But it didn't distract me that much. It was more if you looked behind yourself, you could hear some audience, but you couldn't see anything. You could only see some shadows moving once in a while. I think that it might have added to the experience, if it was more visible.

Participant 17:
I: Which of the experiences (A,B,C) did you prefer, and why?
P17: I'd like to say A, but I think it more sound like reality than the other two sounded like. I feel a was more in it with the sound of A than I was the ones was more "redigated" (edited) or something like that in the other two, it's my opinion.
I: Which one would you then say is the second best?
P17: C, but I can't really explain why I'm not saying B. I just think when I'm trying to listen to both of them, the C was better than B.
I: What did you think of the music that you heard? The genre?
P17: The genre was really good. That kind of genre was a really nice choice instead of heavy metal or anything like that.
I: What do you listen to normally?
P17: All kinds of music, but I would say that is one of the genres I would prefer to go to a concert to – because it's important to.. the staging, as well as the sound, and the whole of it, and just putting it in a CD. If that makes any sense.
I: What did you think of the equipment you had on?`
P17: I was a first timer user, but I think it was really funny to try it – one of the questions was that I felt seasick. I was a little bit, because of my position in the room made me a little bit seasick. That I'm sitting up and looking down at them, but they are still really tall, and then I felt really small. But I felt comfortable wearing all the equipment.

I: Do you have other critical comments?
P17: The one I was thinking about. When you're listening to that kind of music, I'm imagining myself having a beer, and maybe dancing a little bit. And then the equipment will maybe.. When I tried walking I was feeling really locked, because it was really difficult to walk when you don't know where you are. But that's it.

Participant 18:
I: I just realized that I have accidently pressed pause, so just up-sum
I: Which of the experiences (A,B,C) did you prefer, and why?
P18: Yea, what did I say…
I: You said A and C..
P18: … were quite the same, so I don't know if there were any difference in the sound. At first when I sat with B I thought that the quality of the sound was bad.
I: And you normally listen to radiohead?
P18: No just radio hits, not radiohead – you know just mainstream music, also like 80s and pop.
I: Yea, you said high compared to..That the level of the music was too low

P18: Yea, to get an actual experience out of it.

I: Do you have other critical comments?
P18:

Participant 19:
I: Which of the experiences (A,B,C) did you prefer, and why?
P19: A. It was more like I was in the room, and the sound was realistic. And the other videos was more like if I was listening to a song on YouTube. So it felt like I was more at a concert with the sound.
I: Can you describe why it felt like you were there more?
P19: I don't know much about sound, but it just sounded more like someone is singing into a microphone, and I'm like right next to them. And less studio-like. If that makes sense. But I don't know which was the first one I listened to, like before (the quantitative test)
I: I think you started with B
P19: Okay. Because the first time I had my hair down in front of the ears, so I think that did something. But yea A, B and then C. That's how I would rate them.
I: Why do you think B is better than C?
P19: I think C it was like the sound was further away.
I: What did you think of the music that you heard?
P19: I liked it. I liked the part with the guitar in the A and the B video. And I think the C one, maybe I didn't listen to it in that one. But that sound was more like a studio sound or like a CD or.. Cd, hahaha.
I: MP3 maybe, or Spotify song – I know what you mean. What kind of music do you normally listen to?
P19: A little bit of everything. The only thing I don't listen to is country music. And rock n roll like hard core, heavy metal. I'm not that much into that. And country. Those are the only genres I don't really listen to.
I: What do you think of having the headset on, and experiencing the concert?
P19: The first time I tried it on, it was a little loose. But then the second time, the experience was much nicer. It was just like, you know, I was standing on the stage, and I couldn't help wondering if you guys filmed it from the stage, like if you were standing on the stage, or..
I: It was on stage.
I: Yea, the camera was.
P19: But sometimes I kind of wanted to take a step back, just to see the whole band together. But it was fun being on stage too, I liked that part of it. A couple of times I was like uh, now I'd really like to just step back and then view a bit. Yea.
I: That's in version 2.0
[…]

I: Do you have other critical comments?
P19: Not really, other than it was a fun experience. I would definitely try it again.

I = INTERVIEWER
P# = PARTICIPANT

Participant 20:
I: Which of the experiences (A,B,C) did you prefer, and why?
P20: I actually preferred the first video, video A. I think it's because I felt like it was more, somehow, realistic, and it was more like being there. And I think that there was different levels of sounds. Yea.
I: What was you second best?
P20: I think it's definitely video B.
I: Why is B better than c?
P20: I think that C has a more like mono sound, where B had more.. I think the bass level in video B was a lot higher than A, and also C. And I think that made the video better than video C. but still I think video A had the better like "samspil" with the music sound and the vocal sound of the artist.
I: What did you think of the music?
P20: It was actually really nice to hear it.
I: What kind of music do you normally listen to?
P20: Underground music, RnB, stuff like that.
I: What did you think of having the headset on?
P20: I don't know. You kind of get in a situation, and a context, when you get the earphones on. But I don't think it was negative. You kind of zoomed out and got in another context than you were in this room.
I: What about the navigation and all that?
P20: It was kind of fun in the start, because you're looking at something that isn't there. But it was cool because you only see particular space, and now you can move your head and see other thing else. It was a fun experience.

I: Do you have other critical comments?
P20: I don't know about the quality of the video. I think it was kind of "grynet", but maybe that's just because you can't make HD videos with the virtual stuff. Maybe you can look on that. Maybe easier than done.

Participant 21:
I: Which of the experiences (A,B,C) did you prefer, and why?
P21: A. The sound is louder in A. And – jeg er ikke så god til engelsk.
I: Så kør den på dansk, hvis du hellere vil det.
P21: Jeg synes bare lyden lød mere tidligere, og den var højere i A'eren.
I: Hvad var så næstbedst?
P21: Jeg synes B og C var meget ens. Jeg kunne ikke høre så meget forskel. Men jeg synes B'eren var lidt mere tydelig end C'eren. Så det er A, B, C.
I: Hvad synes du så om musikken?
P21: Altså, om det er godt eller dårligt? Det er ikke noget musik jeg høre.
I: Hvad hører du normalt, da?

P21: Det er rigtig forskelligt. Men jeg hører meget arabisk musik. Jeg er selv araber, så det er ikke fordi jeg hører sådan til det danske eller engelske.
I: hvordan synes du det var med med det her headset?
P21: Det er første gang jeg prøver det, men jeg har set det før, hvor man kommer ind i en helt anden verden. Jeg synes det er rigtig fedt. Jeg tror bare den eneste ulempe er, at de filmer jo sikker fra siden af, og det er ikke foran så man får det hele med, så man skal kigge rundt. Hvor at, for mig, har jeg større overblik når jeg kan se det hele på engang, i stedet for at kigge forskellige steder.

I: Do you have other critical comments?
P21: Nej altså, kvaliteten kunne godt være bedre, nu hvor man er vant til HD tv. Man er vant til meget bedre kvalitet. Men jeg synes det er rigtig godt alligevel.

Participant 22:
I: Which of the experiences (A,B,C) did you prefer, and why?
P22: C. The sound was better in C.
I: Can you elaborate why it was better?
P22: Jeg er ikke så god til engelsk.
I: Så tager vi den bare på dansk.
P22: Jeg føler bare at der var mere kvalitet i lyden.
I: Hvad var så den næstbedste?
P22: Det var B. Der kunne jeg ikke høre så meget på højre side. Der var lyden lidt svagere.
I: Så hvis du skal ranke dem som bedste, næstbedste og sidste?
P22: C, A, B.
I: Hvad synes du så om selve musikken?
P22: Det er da god musik, det er fint nok.
I: Hvad for noget musik lytter du mest til?
P22: Det er lidt forskelligt. Altså, jeg er kurder så jeg hører meget kurdisk musik.
I: Hvordan synes du så det var med headsettet, at have det på?
P22: Jeg har ikke prøvet noget andet, så det var fint nok som det var.
I: Hvad med navigeringen af det?
P22: Det var selvfølgelig lidt svært i starten, men man lære det hurtigt.

I: Do you have other critical comments?
P22: Næ, ikke rigtigt.

Participant 23:
I: Which of the experiences (A,B,C) did you prefer, and why?
P23: I think I prefer number A. Because it felt more realistic if you're at a concert. It was more.. What's the name, spatial? I'm not a music nerd, so I thought it was a bit tough, but it reminded me the most of a real concert, I think. [laughing]
I: What was the second best then?
P23: Oh, I'm not really sure. I think they reminded a lot of each other. Maybe B. No.. I don't remember. But yea. I will say B.

I = INTERVIEWER
P# = PARTICIPANT

I: I wanted to ask you why, but if you.. yea. A is the best.
P23: Yea [laughing]
I: What did you think of the music that you heard?
P23: Yea it was catchy. Not my favorite music, but she was pretty cool, the singer. Actually, the more I heard, the more I could feel my body like rocking. So it definitely catched me at some point.
I: what did you think of having this equipment on, and experiencing through that?
P23: It felt real. It felt like you were there. But in the beginning I thought it was weird, the angle, because I had more attention at him because it was a better angle, it was the bass player - but then at some point in the movie she moved, and then you could really see her. So that's the only thing, it would be really cool if you could also move. But yea, that's the next step. And I'm a big fan of VR. But you can feel it's heavy on your head. And you can get a bit dizzy. I have tried other Vr glasses, where I got really dizzy. That didn't happen here, and I was also siting down. But maybe if I had them on for like an hour I wouldn't be able to, yea.

I: Do you have other critical comments?
P23: No, I think it was pretty cool. It felt like you were there. And actually I also think I wrote it in the (questionnaire). That at some point the music was actually better than if you were at a real concert, because the speakers can be really bad at a real concert. Especially when it's a small place like this. So at some point I was like yea maybe this is the way to do it. If you have really good speakers that could be interesting.

Participant 24:
I: Which of the experiences (A,B,C) did you prefer, and why?
P24: I preferred number A, actually, after I heard the others. Because I kind of get the feeling that I like the most realistic one. And I thought that number A was the most realistic if you're at a concert. But maybe the quality I thought maybe B. But I thought that A was more realistic and more concerty.
I: So you prefer A, and then what is your second choice?
P24: B.
I: and then..?
P24: C.
I: And why is B better than C you think?
P24: Maybe the quality, but I'm not an audio person. But yea, I thought the bass was better, I think. You know what the real answer is maybe.
I: There is no real answer, or wrong actually, it's mostly what people actually think and how the experience it.
I: What did you think of the music?
P24: It was fine, funky. Yea. I thought it was quite allright.
I: What kind of music do you normally listen to?
P24: A lot of different genres. But maybe not that type of music. Yea, what do I listen to? All genres. But not rock and metal. But hiphop, pop, RnB.
I: What did you think of this headset and the experience with it?

P24: I thought it was quite all right. Those things that were sitting on your nose was a bit. You didn't maybe feel like you were in the room that much, because you could feel that it was on your face. But when I looked around, I felt that I was in the room because of the spatiality was kind of good. But I haven't experiences VR before, so that's maybe why I think it's a very nice experience.

I: Do you have other critical comments?
P24: No I think that was it.

Participant 25:
I: Which of the experiences (A,B,C) did you prefer, and why?
P25: The A. The sound is more clear and realistic. And it is really important because you feel like you were there.
I: Which one is the second one then?
P25: B,
I: Why?
P25: It's a little more better than C, which is really bad.
I: Why do you think C is bad?
P25: Cannot explain. The way you hear the song. It was really low first of all. I couldn't hear the details, you know.

I: Do you have other critical comments?
P25

Participant 26:
I: Which of the experiences (A,B,C) did you prefer, and why?
P26: It's actually either the first or the third. Cause I'm not sure the second was implemented well. I felt it would have sounded differently.
I: By first and second, you mean first by A, and second B…
P26: Mhm, the first where it's audio from listeners perspective. The second one where it's 360 dome sound perspective, also from listener actually, and the third one where it's directly from mics and all perspective. Studio-like, isn't it?
So, for me I preferred where it didn't change. But perhaps it's just because the second one. Something didn't feel right for me, at some angles when you were the sound would be really dim as if you were almost out of the venue. And that didn't feel believable to me.
I: So the one you preferred were A?
P26: Mhm
I: and what's your second choice?
P26: It's C.
I: And then B as the last. Cool. […] What did you think of the music for this music?
P26: I do like it yea.
I: What kind of music do you normally listen to?
P26: Anything what's not being played on radio stations?[…] Well, I do realize that live concerts provide way different experience, and this was groovy too, which felt enjoyable. And this wasn't

just a live recording which you usually listen. This was attempting to put you in to a situation where you are closer to the performer. I enjoyed it. But I also have other comments about the position of where you are later.
I: I was going to ask you now. Or actually, I was going to ask you also how it was with the headset, and experiencing a concert with it.
P26: It's all well. I guess you could have mentioned that you can actually fast-forward videos.
I: I'm sorry I forgot to say that.
P26: I was wishing to do that, and then accidently found it [laughing]. Well, experience, I guess I would love to be able to choose different spots. 'Cause for me it felt like I'm a little too close. I even watched the videos of Circe de soleie – they have some circuit performances where you're also on the stage, and everything is happen there. And I'm personally not fan of this part of the camera being this close to the action. If I wanna be a piece of audience perhaps it should be a little back. And therefore it would be great to choose several spots. Far back, closer, really close.

I: Do you have other critical comments?
P26: I don't think so.

Participant 27:
I: Which of the experiences (A,B,C) did you prefer, and why?
P27: I think either.. I think if we have to consider it in a concert-setting, it would be either A or C. I really liked the effect that you get from when you turn the head around, that you just listen, and get the sense of the instrument that your ear is directly towards if you turn your head around. But for concert-wise I would probably think that A is the most suitable as something you would be able to relate a bit more to. Whereas I think that I got the whole sense of like everything in the music, so I wouldn't have to be afraid to turn my head around or trying to get every detail in the track.
I: Okay, and that was for A?
P27: That was for C. The A thing again, that was something I probably would relate, for me personally, would relate more to the whole concert setting-ish. Where depending on where your head would turn, that would give a notion of what instrument you would hear than the other.
I: So you prefer A, and the second place would be?
P27: That would be C.
I: and the third place B?
P27: Yea.
I: So why is C better than the B one?
P27: I am not very fan of that whole muffled, yea the muffled sound when I turned away. That distracted me a lot from the music itself. I was more focusing on the effect that was happening than the music piece.
I: What did you think of the music that you heard?
P27: I think it was very catchy, I really wanted to keep listening to it.
I: That's good – Jacky Venson, if you want to look it up. How about the headset equipment?
P27: What do you mean?
I: How was it experiencing a concert through a headset?

I = INTERVIEWER
P# = PARTICIPANT

P27: It was definitely a new experience. I have never tried it before. I would say I much prefer going out to a regular concert, but I definatly think that it gives a much better addition than if you were to watch it regularly on the internet, that I definitely would prefer this one instead, because you feel a bit more immersed in the sense that you can actually see the band, look around, see where they are instead of having the static image shown to you.

I: Do you have other critical comments?
P27: Uhm, not really. I actually do like that you cannot see the audience. I was wondering where the audience was, you could not see them clearly. But I think that it would draw too much focus away from the actual band. It allowed me to have more focus on the actual players on the band and listen to the music, instead of turning my head around constantly.