Super hearing: a study on virtual prototyping for hearables and hearing aids

Master Thesis

Luis Vieira

Aalborg University
Department of Architecture, Design & Media Technology
M.Sc. Sound and Music Computing

Department of Architecture, Design & Media Technology
M.Sc. Sound and Music Computing
Aalborg University
http://www.aau.dk

AALBORG UNIVERSITY
STUDENT REPORT

Title:
Super hearing: a study o virtual prototyping for hearables abd heading aids

Theme:
3D Audio, Virtual Reality, Hearing Aids, Virtual Prototyping

Project Period:
Spring Semester 2018

Project Group:
Master Thesis

Participant(s):
Luis Vieira

Supervisor(s):
Stefania Serafin
Michele Geronazzo
Jesper Udesen

Copies: 1

Page Numbers: 56

Date of Completion:
August 31, 2018

Abstract:

Directionality in microphone array systems for artificial hearing or hearable devices exposes the user to certain limitations in the way it can perceive the auditory environment - not always the source of interest is positioned directly in front of the user, or not always it is desirable to have one single point of directionality. Attaining for a possible separation of the audio stream into multiple singletons, this project developed three modalities of sonic interaction, to study the benefits of virtual prototyping using virtual reality environments. Seventeen participants took part in an experiment to find what level of precision was demonstrated when subjects used the interaction models. There were statistical significant differences between participants performances, when performing a selective auditory attention voice-pairing task, with the three interactions in a purely virtual rendered scenario and in comparing the interactions' scores between both scenarios.

Contents

Preface

Aalborg University, August 31, 2018

Author 1
<lvieir16@student.aau.dk>

Acknowledgements

I would like to thank my supervisors Stefania Serafin, Michele Geronazzo and Jesper Udesen for guiding me through all stages of this thesis, sharing their knowledge as well as experience and providing me with great opportunities and contacts to apply the outcome of this work. I would also like to thank you DPA Microphones for supporting this project with required equipment and to the Aalborg University Copenhagen which made available the research facilities for me to work during these past seven months.

To all my friends who supported me and motivated me and to those who just like me have been working intensely on their own share of knowledge to the community,and with whom I had the chance to share ideas with throughout my project, I'm grateful and happy to reach the end of this chapter - if a good reward comes along, even better!

To my family, I can only be grateful for their support and to let them know what an amazing and rich experience they allow me to discover.

Thank you and now let's continue!

Chapter 1

Introduction

> There are domains beyond the reach of language, where it is insufficient, where semiotic-conceptual work has to be and is done by means of other modes.
>
> Gunther Kress [27]

1.1 Problem Statement

The human listening experience can be defined by the ability to focus attention in a certain direction, towards a specific source of sound propagation [6]. This selective auditory attention relies itself in a set of mechanisms that work together to help the human brain understand the direction of an incoming sound, as well as what physical features characterize it source, to create in the end an auditory representation of it [6] [40]. In this way, these systems filter the sound field in what is relevant content and what is noise, reinforcing the auditory system. Listening becomes a multi-modal experience, where not only the ears translate the outer sonic reality but also cooperate with bodily and cognitive mechanisms.

In this research, it was important to understand to what degree these mechanisms - body motion, visual feedback, spatial directionality - influence the listening process, how they support the auditory behavior and to what extent can they be used to support artificial hearing devices such as hearing aids and hearables. The final goal defined the ability to control these interactions to create a *super-hearing* set of tools.

In collaboration with the Danish audio company GN Resound, the project was framed as a platform for virtual prototyping of hearing aids, where multiple hearing models could be tested in different situations and with the reinforced use of three interaction metaphors for ideal beamforming control: the head rotation, the eye movement and the hand remote controller. For this framework, two different scenes were proposed to test the user's control with and without the presence of

real people. A first scenario with eight virtual cubes displayed in a perfect circle around the participant and a second one with eight real people displaced in the same position as before, recorded in a 360º video. The participants could operate these metaphors with three degrees of freedom: the direction of a high directivity system, the width of the directivity beam and the sound pressure level of the stimuli. The goal was to understand how efficient could they be with these interactions and these parameters when trying to find multiple pairs of voices in a multi-talker environment, what values they used for these parameters, what strategies to solve the task, body motion and dynamics (between head rotation, eye movement, and body position).

1.2 Research Question

The complexity of this subject might be apparent at a first glance but it's indeed even more demanding when trying to frame a single research path.

Imagine the following scenario: you find yourself in a coffee place, sited at a table with a few friends, ready to order some food and drinks. While waiting for the waiter to come and take your order, you talk about different subjects. It's lunch time, so the place gets slowly busier and busier (and noisier and noisier). However, you and your friends are able to continue the conversation with a bit more effort but still in a very natural way – where you intuitively adapt your body position/language and your speech level to make yourself clear. This ability to ignore the increase of background noise while following a conversation is what is famously called the cocktail party problem and is one of the most powerful abilities of the human auditory system that allows to distinguish between meaningful and non-meaningful auditory information (see figure 1.1). This is also one the main challenges for people with hearing aids. [40]

This study is built on the assumption that even though the current reality of technology doesn't allow a perfect separation between relevant signals and background noise, it can be foreseen that future developments will enhance this feature with, for example, the use of artificial intelligence to analyze the sound field and separate individual talkers in a cocktail party situation. In this sense, the research question can relate to how singled audio streams can be presented to the listener in the most optimal way for a specific task: concurrent multi-speaker.

The process of developing an experiment that could introduce different types of sonic interaction - "by using sound in an embodied and performative way (...)" [46] to affect the perception of the reality or even affect the transmitting message or source,revealed that there can be an ultimate research question if indeed no real constrains are simulated, if ideal parameters and extreme conditions for listening experiences could be attained. This research question would investigate the potential use of virtual prototyping in testing multiple listening conditions with different

interaction metaphors. A first level of questioning can see how the user will experience a listening situation in a realistic environment with such ideal systems. It is thus necessary to define and evaluate this perceptual experience and frame the research question into different aspects:

- Which interaction would help the most when auditory attention needs to be reinforced? Is a more embodied interaction a more efficient interaction?
- When given specific parameters of control over the stimuli, are there optimal combinations of these parameters that can be generalized?
- Any individual differences and user profiling

Virtual reality was chosen as the platform for this experiment has it provided more flexibility and faster prototyping of these different sonic interactions as well as the parameters of control over different setups of experiment, in this case, a complete virtual environment and a 360° video rendered environment, providing different levels of abstraction in the environments: from minimal visual feedback to realistic recordings.

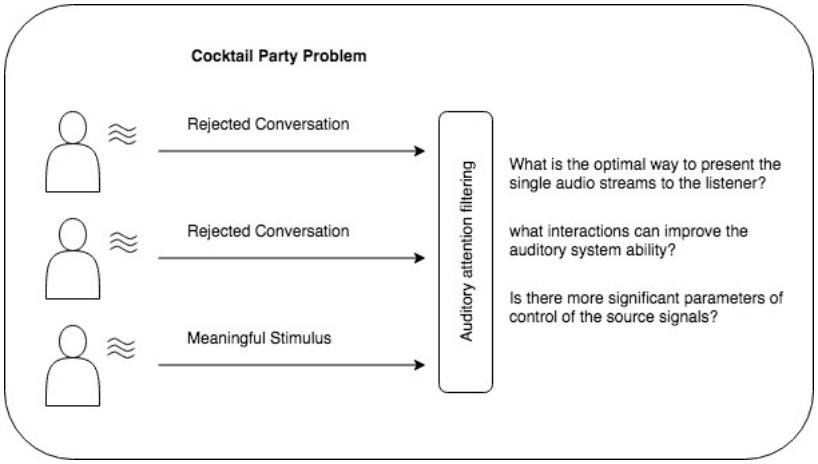


Figure 1.1: Cocktail party problem described in the scope of this project.

Hence, the sonic interactions developed were designed to reinforce the listener's selective auditory attention through the control of a directivity beam that could select a specific speaker, attenuating the presence of unwanted sounds.

1.3 Outline of the report

The report is organized in 4 main parts:

Chapter 2 is a three-fold chapter exploring some of the theoretical background that supports listening cognitive processes, listening behaviors as well as how to frame listening as a information processing system so that it can be possible to look at the experiment results and find any data that could be related to how we listen, how we focus attention, and what physical and motor strategies do we use. It extends this knowledge to the science behind beamforming and how directivity patterns influence the spatial information received by the listener. The second and third parts relate to the spatial audio rendering - the basic theory behind spherical encoding of virtual sources, and the extension to higher orders of ambisonics (HOA) and the implementation behind the ambisonic software used (Resonance Audio).

Chapter 3 defines the material and methods, i.e. technical implementations required to build the experiment virtual environments, the software and hardware used for the virtual reality rendering, and how the sonic interactions were implemented. On the second half of the chapter, the experiment protocol is described in detail.

Chapter 4 reviews the data collected, and in what perspectives it was analyzed, providing some statistical results between conditions of interaction, and conditions in the two experiments.

Chapter 5 and **Chapter 6** review the work developed, what conclusions can be drawn and what meaningful knowledge can be taken from these experiments and what future works could provide, as well as some of the participants comments are shared.

Chapter 2

Fundamentals

2.1 The multi-dimension construct of listening

> Consists of four connected activities – sensing, interpreting, evaluating, and responding.
>
> *Steil et al. 1983*

2.1.1 Aspects of cognitive and information listening processing

In order to contextualize the current study, this section revises some of the literature, the theories and the technologies developed around listening behaviour and artificial hearing systems.

The cognitive aspect of how the human brain chooses to selectively attend to certain characteristics of the surrounding environment, processing it in a high level of detail, while simultaneously being able to reject other stimuli that are regarded as less relevant to interact with, is a fundamental question to further understand what defines the way auditory information is taken and processed.

Listening became a relevant part of human selective attention research since 1953. Starting with Cherry's research, the "cocktail party problem" [9] and the use of dichotic stimuli to test speech intelligibility. It was since then recognized the brain's ability to process incoming information at different levels of perception, but also as an acoustic phenomena, masking (the effect of interfering signals in speech recognition) and binaural processing. All these factors contribute to our ability to segregate signals – also referred as auditory signal analysis. [4] [6]

Communication under the "cocktail party" conditions is one of the most powerful skills of the human auditory system. When confronted with multiple simultaneous stimuli (speech or non-linguistic stimuli), it's necessary to perceptually

segregate relevant auditory information from concurrent background sound and to focus attention on the source of interest [9] , [6]. This segregation is related with the principles of auditory scene analysis [4] in which a stream of auditory information is filtered and grouped into a number of perceptually distinct and coherent auditory objects. In multi-talker situations, studies on spoken language processing suggested that auditory object formation, object selection, and attentional allocation – the ability to focus on an auditory object of interest - are closely related to each other and can be described within a model of successful "cocktail-party" listening. [20], [45]. Listeners use different strategies to solve the "cocktail-party" problem. These strategies are important in dynamic auditory environments, where changing between auditory objects requires more effort in scene analysis and selective attention. [2]

In 1958, Broadbent's filter theory [5] laid an important connection between the psychological phenomena and information processing concepts of multi-stimuli listening situations. His theory depicted cognition as a series of discrete, serial information-processing with a two-stage processing of the attentional limits. A first stage regarding the physical properties (such as pitch and location of the sound) that would be extracted for all incoming stimuli, in a parallel manner; And a second stage that required more complex psychological properties, that would go beyond the physical characteristics (e.g. the identity or meaning of spoken words). In this stage, the processing ability is more limited and thus a selective filtering would protect the system to overload, passing only those stimuli which had a particular physical property, from among those already extracted for all stimuli within the first stage. This model was coherent with Cherry's conclusions on the selective shadowing/masking ability of the human brain to ignore less meaningful content, structuring a base for further understanding of the human ability to listen.

Cognitive research has shown that listening comprehension is more than extracting meaning from incoming speech. It is also a process of matching speech with what listeners already know about the topic. Therefore, when listeners know the context of a text or an utterance, the process is considerably facilitated since listeners can activate prior knowledge and make the appropriate adjustments to an understanding of the message. [8]

There can be considered two distinctive processes involved in listening comprehension: a "top-down" approach which uses prior knowledge to understand the meaning of a message, the topic, the listening context, the text-type, the culture or other information stored in long-term memory as schemata (typical sequences or common situations around which world knowledge is organized) and a "bottom-up" approach when the listener uses linguistic knowledge to understand the meaning of a message, building meaning from lower level sounds to words to grammatical relationships to lexical meanings in order to arrive at the final message. Listening is in the end an interactive, interpretive process where both approaches –

prior context and linguistic knowledge – work together to understand the stimuli. These processes are in their own way influenced by the listener's knowledge of the language, familiarity with the topic or purpose of listening. [49]

Dichotic listening and masking effect

Dichotic listening, related with the hemispheric lateralization of speech sound perception [21], is a psychological test used to investigate selective attention in the auditory system. It shows the brain's ability for hemispheric lateralization of speech perception, a feature of relevant importance when listening to different acoustic events presented to each ear simultaneously [34]. While performing the dichotic listening test, the participants are exposed with two different stimuli simultaneously which are directed to each ear independently. The task resolves on the ability of the participant to focus attention to one or both stimuli and consequently described the content of the message they were asked to pay attention to or the one they were supposed to ignore. Some further research shows the participants' capacity to adopt to what type of response is necessary and thus adapting the dichotic listening to the context. Musiek and Pinheiro [16] reported two common conditions: the listener's skill to repeat stimuli directed to both ears, which requires binaural integration – also referred to as the free recall/divided attention response mode - and a second common response that regards the binaural separation, when the listener needs to focus on the stimuli presented only to one of the ears and ignore the stimuli presented to the opposite ear – also referred to as directed attention or directed report. Related to this ability, the right-ear advantage (REA) is perhaps the most interesting finding [25] revealing the direct anatomic connection of the right ear to the left hemisphere, which in most people is specialized in language processing.

Going back to the cocktail party scenario, there is another condition that can affect the efficiency of dichotic listening: the effect of interfering speech or other non-linguistic signals. These are conditioned by the frequency spectrum characteristics of the signals and their spatial information. Byrne et al. [7] demonstrated that the long-term average spectrum of speech is constant across 12 languages. There is a systematic difference between male and female speech, which only occurs at the low end of the spectrum (around 100 Hz), caused by the fundamental frequency of the male voice, which sinks at around 90 Hz in contrast with the female voice that sinks around 150 Hz. The average spectrum peaks at 500 Hz, falling off with a slope of 6 dB per octave until it levels at 4 kHz. These differences can be accentuated when speech is whispered, shouted or spoken. [11]

Associated with the spatial information of the interfering signals, the human anatomy of the body, the head and the ears reinforce the physical characteristics of the stimuli and helps the brain understand where to locate the sound in space. Mills [32] found a 1 degree of audible angle resolution when a sound source is

located straight ahead. The interaural time differences (ITD) and interaural level differences (ILD) between both ears work together to improve the release from masking effect - when the signal of interest shares the same spectrum identity and/or the same sound pressure level with interfering signals – and increase the decorrelation between left and right ear and thus the separation between background noise and meaningful sounds. Beyond these two factors, the brain also correlates the signals that arrive at both ears, also known as IACC, which is a measure associated with the feeling of of spaciousness and envelopment in acoustics and the higher this value the more spacious and comfortable the space feels for the listener. [] The information embodied in interaural time differences (ITDs) and interaural level differences (ILDs) allows listeners to locate sound sources on the horizontal plane and have an important role in generating high levels of speech recognition in complex listening environments part.[31] ([3]; [9]; [55]; [47]). ILDs are dominant for signals with frequencies above 1500 Hz since the wavelengths are short compared to the head dimensions. Our threshold for detecting ILD is about 1 dB. ITDs are dominant for frequencies below 1000 Hz, low to mid frequencies, where the auditory system can sense this time difference by comparing the phase of the two ear signals [3],[47]. The threshold of ITD detection is about 10 μs [26], which corresponds to a frequency of 100 kHz, far above the audible frequency range. This is the reason for the 1-degree localization accuracy when the source is located directly in front of the head. Previous studies have focused largely on the use of ITD information to separate speech from background noise. There is little information about whether ILD cues provide this same benefit.

Also, very remarkable in speech levels and intelligibility is the Lombard effect – the tendency of speakers to increase their vocal effort in the presence of background noise. From studies by Lane and Tranel [29], it can be concluded that the effect amounts to 0.5 dB for each dB increase of the noise level above approximately 50 dB SPL. [29] With this in mind, it seems that signal-to-noise ratio (S/N) is a better base of analysis of acoustics of interfering speech and speech intelligibility. Speaker and listener will naturally adjust vocal effort, distance and head orientation to maintain this ratio as best required. [6]

Researchers have addressed a wide range of factors that influence the speech intelligibility in the presence of competing speech and some methods of evaluating this relation have been provided to measure the listeners' ability to identify and understand the receiving message. There are two particular interesting topics of research defining the causes and methods associated and used by listeners to naturally maintain the aspect of signal-to-noise ratio at the right value for communication and listening experience.

The first of these topics is the interfering effect of competing speech (normally in a complex listening scenario like the cocktail party problem), which causes the

excess of masking and in the case of non-linguistic signal, informational masking. Associated with this excess of masking, listener's will perform physical efforts to accomplish voice segregation necessary, such as the binaural listening, invoking the spatial separation of the sound sources, the ITD and ILD factors associated with the head shadowing. Culling and Summerfiled [12] shown that listeners were unable to use ITD for segregation of concurrent synthetic vowels which suggests that the binaural gain was caused mainly by head shadow.

The second topic relates to the influence of the cognitive aspect of listening, auditory attention and the perceptual mechanism of selective listening. Several studies proven that to solve the masking produced by normal speech or other stimuli, no meaningful importance was given to the content of the interfering speech, and that listeners are capable to actively follow the target speech while almost completely ignore a secondary speaker. It is clear that auditory information is processed in both parallel and sequential order of input management. Wood and Cowan [54] demonstrate that listeners who note specific changes in the "unattended" channel perform poorer on their main task (shadowing the target speech), which indicates that part of the processing becomes sequential. This would mean that performance will be better for selective rather than divided-attention tasks. We can then refer to speech segregation and attention as higher and complex information processing in the listening experience that are correlated with the relevant effects of masking and binaural unmasking. [6]

For people with hearing impairment, understanding speech in noise situations becomes a very difficult task when both speech and noise co-exist above their hearing threshold, the ability to focus attention only on the important stimulus is affected and they are not able to release the meaningful message from the mask as well as benefit from the necessary signal-to-noise ratio (SNR) for an optimal intelligibility.

2.1.2 Hearing impairment, hearing aids and methods of analysis

Typical hearing losses are located in the cochlea where hair cell damages can be observed, often provoked by loud sound exposure. [38] Hearing losses generate large problems of communication of the affected persons, who find it much more difficult to understand speech in noisy environments, particularly with multi-speaker scenarios. There are two effects that can be described as consequence of these impairments: An increased hearing threshold, also known as SRT (speech reception threshold), describing the lowest level at which a person can separate meaningful signal from noise. Normally this value ranges from a few dB to more than 10 dB causing severe problems of communication. Hearing capacity can be described by the area between the hearing threshold and the uncomfortable level that characterizes a pain threshold. (sound pressure vs frequency) [38] (see diagram 2.1) When in the presence of hearing impairment, this area is defined by an increase of the

hearing threshold (SRT) and a decrease of the uncomfortable level.

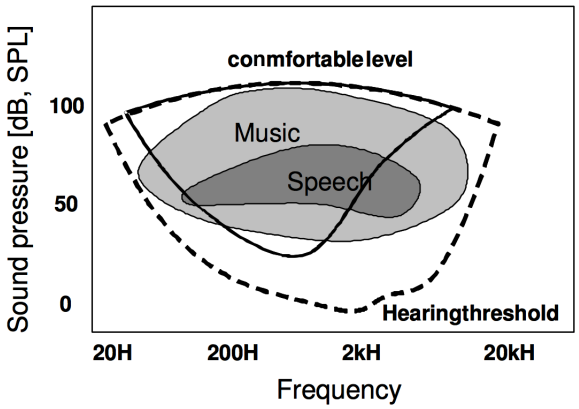


Figure 2.1: Hearing threshold and uncomfortable level of normal(dashed line) and hearing impaired (solid line) persons.

The other relevant impact of the hearing impairment is the stronger signal masking. Due to the damage of the outer hair cells in the ear, the resonance effect and corresponding frequency perception are reduced. Because of this, hearing impaired people notice a severe constrain on their ability of speech intelligibility and a signal mixture of noise and meaningful signal. (see figure 2.2) The brain is no longer able to benefit from various factors such as the long-term spectrum fluctuations, where speech is recognized to have larger variance, helping the normal hearing identify and follow these signals (with a gain of 7 dB) in comparison to 0-2 dB for impaired hearing [6]. Several studies [38], [18], [6] show the impact on the masking release, reduced even in spatially separated sources, particularly for the head-shadow component since it occurs at high frequencies where the bigger loss happens.

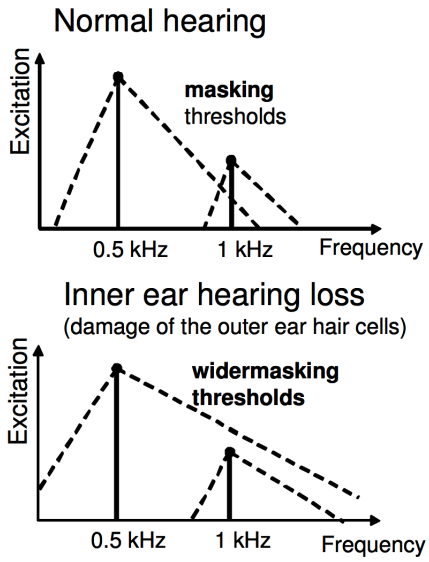


Figure 2.2: Hearing threshold and uncomfortable level of normal(dashed line) and hearing impaired (solid line) persons.

In the figure 2.2, there's an example of the strong masking caused by hearing loss, where the frequency 1 kHz is no longer detected.

Hearing aids aim at compensating for these two major effects in the hearing impairment. There are two standardize types of hearing aids: Behind-The-Ear (BTE) and In-The-Ear (ITE) devices. BTE devices generally allow for the compensation of stronger hearing losses. [38] More recently, alternative designs came to the market, smaller in size and with a thinner sound tube that connects the hearing device behind the ear to the ear canal, called Receiver-In-Canal (RIC). These allow for a more comfortable wear but also reduce the occlusion effect, where the users' own voice sounds unnaturally dull.

The requirements of signal processing for hearing aids are very restricted and concern the signal delay, complexity of the processing system and the beamformer restrictions due to the physical size of the device and optimization due to energy consumption. In general, the signal flow (see figure 2.3 starts by capturing the acoustic input with a microphone array that can be constituted up to 3 microphones processed into a single signal within the directional microphone unity. To compensate for the reduced hearing area a frequency dependent compression is applied by an analysis filterbank and a corresponding signal synthesis. The main frequency-band-dependent processing steps are noise reduction and signal amplification combined with dynamic compression. Indirect methods address the problem of strong masking and try to increase the SNR of the signal output of the hearing aids by beamforming or other noise reduction approaches.[18]

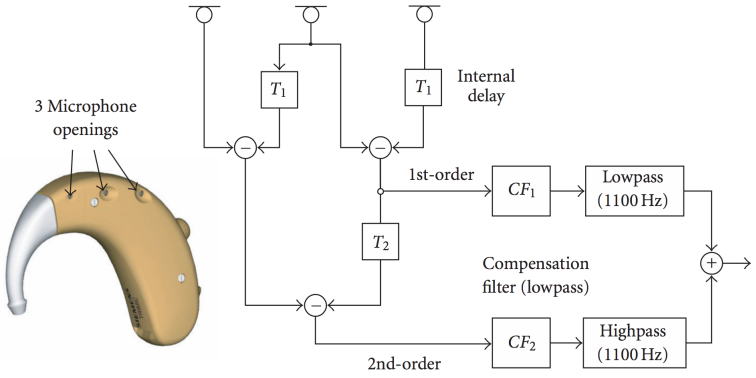


Figure 2.3: Artificial hearing system (BE) model and the DSP diagram of the device. Taken from [40]

Directional microphones, adaptive direction microphones (beamformers), binaural noise reduction

> One of the main problems for the hearing impaired is the reduction of speech intelligibility in noisy environments, which is mainly caused by the loss of temporal and spectral resolution in the auditory processing of the impaired ear. [18]

To compensate the SNR, estimated to be around 4-10 dB for hearing impaired [15], [40], and to help the natural directivity of the outer ear, directional microphones have been used and proven to increase speech intelligibility, and the speech reception threshold (SRT) in the range from 2 to 4 dB [48]. Currently according to the number of microphones built in the hearing aid, there are two main differential arrays: first-order and second-order differential arrays. (see figure 2.3)

In a first-order differential array, directivity is a product of differential processing of two nearby omnidirectional microphones in endfire geometry to create a direction-dependent sensitivity. The signal in the rear microphone is delayed and subtracted from the signal picked up by the front microphone and the directivity pattern of the system is defined by the ratio r of the internal delay T_i and the external delay due to the distance between the two microphones – normally between 7 to 16 mm.

$$P(\Theta) = |\tau/T + cos/\Theta| \tag{2.1}$$

where T is the ratio between the physical distance between the two microphones and the speed of sound.

To compensate for the high-pass characteristics introduced by the differential processing, a low-pass filter is usually added to the system. The directivity Index

(DI) is a measurement of the performance of a directional microphone, defining the power ratio of the output signal (in dB) between sound incidence from the front and the diffuse case – from sound coming equally from all directions. DI can then be interpreted as the improvement in SNR that can be achieved for frontal target sources in a diffuse noise field. Regarding the performance related to speech intelligibility, a weighted average of the DI across frequency is measured, also referred to as AI-DI. The weighting function is the important function used in the articulation index (AI) method [33] and takes into consideration that SNR improvements in different frequency bands that contribute differently to the speech intelligibility.

The second-order array can be understood as an improvement on the two-microphone array setup previously described. The reason for the introduction of a new microphone in the array is due to the high sensitivity to microphone noise in the low frequency range. The third microphone is implemented in series with a high-pass filter limiting the processing to the frequencies above 1 kHz – also more relevant for speech intelligibility. This extra processing improves the AI-DI up to 2 dB compared to the first-order system, with values of 6.2 dB.

> For many listening situations, improvements of 2 dB in the AI-DI can have a significant impact on speech understanding [40].

If the desired signal and interferes occupy the same temporal frequency band, then temporal filtering cannot be used to separate signal from interference. However, the desired and interfering signals usually originate from different spatial locations. This spatial separation can be exploited to separate signal from interference using a spatial filter at the receiver. [40] Implementing a temporal filter requires processing of data collected over a temporal aperture. Similarly, implementing a spatial filter requires processing of data collected over a spatial aperture. Beamforming processing technique is an important technology that by employing an array of transducers can increase the receiver sensitivity in the focused direction by decreasing the sensitivity in the directions of interference or noise.

Beamforming algorithms may be categorized into fixed and adaptive beamforming [50]. Fixed beamformers have a fixed spatial directivity (not dependent on the acoustical environment), and focus on a wanted sound source, thereby reducing the influence of background noise, more precisely to attenuate signals outside the line of sight. Examples of fixed beamforming are delay-and-sum beamforming [22], [10], weighted-sum beamforming (Gallaudet and de Moustier, 2000), superdirective beamforming [23], and frequency-invariant beamforming [52].

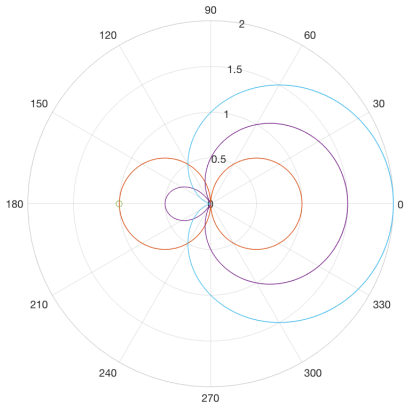


Figure 2.4: Directivity patterns from a perfect bipolar pattern with a digital delay of 0s to a cardioid pattern, where the delay is equal to T_i

In the case of adaptive beamforming, directivity is dependent on the acoustical environment in which the beamformer is located. In high-end hearing aids, the directivity is normally adaptive in order to achieve a higher noise suppression effect in coherent noise, that is, in situations with one dominant noise source [40] [37]. The direction from which the noise arrives is continually estimated and the directivity pattern is automatically adjusted so that the directivity notch matches the main direction of noise arrival. The steering of the directional notch has to be reliable and accurate and should not introduce artifacts or perceivable changes in the frequency response for the zero-degree target direction. The adaptation process must be fast enough (< 100 milliseconds) to compensate for head movements and to track moving sources in common listening situations, such as conversation in a cafe with interfering noise. In figure 2.5 a comparison of the same measurement for a non-adaptive supercardioid directional microphone (solid line) and an adaptive one (dashed line) shows higher suppression effect for noise incidence from the back hemisphere is clearly visible.

Methods of measurement and intelligibility analysis

There are two general categories of methods for assessing the SNR advantage provided by directional instruments: electro-acoustic and behavioral evaluation.

The general term directivity is commonly used to describe electroacoustic evaluation of directional properties. To evaluate the user experience aspect of these technologies, the term directional benefit can be used to describe situations in which a person using a directional mode performs better than when using an omnidirectional mode. Directional research across hearing aid brands sometimes reveals little correlation between listeners' relative performance with directional hearing aids and directional benefit [41],[42],[43]. Performance (absolute score) is

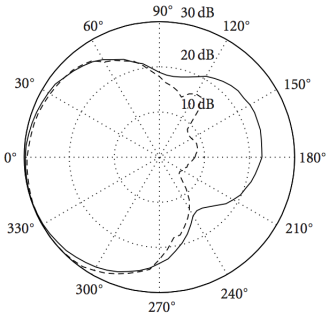


Figure 2.5: Suppression of a noise source moving around the KEMAR for a BTE instrument.

influenced by the hearing aid as a whole, including not just the directional microphone, but also all other signal processing and frequency shaping properties. We can that assume that it relates with the technical specifications and performance of the equipment. In contrast, it is assumed that directional benefit (difference score) reflects the impact of the directional microphone on the hearing aid processing system. That is, directional benefit is assumed to mainly reflect differences in the electroacoustically measured directivity of directional and omnidirectional instruments [41] [43]. We can than relate this value with the quality of the directional microphone behavior.

The three most common metrics of evaluating directivity of a system with hearing aids include the Front-to-back Ratio (FBR), directional patterns and the DI [40]. FBR is mainly used in clinical procedures and directional patterns and DI relate to the use of anechoic environments. Directional patterns of hearing aids are commonly measured in a single (horizontal) plane and are graphically realized by a two-dimensional polar coordinate system, providing an easy read and comparison of directivity behavior of different hearing aids. The magnitude of relative hearing aid output is plotted as a function of the distance from the center of the sphere. That is, a smaller sphere is reflective of greater average attenuation. In addition, the angles of greatest attenuation, usually referred to as nulls, are displayed as indentations in the sphere.

While directional patterns can provide detailed information relative to the attenuation provided by a hearing aid across angles, it is sometimes difficult to visualize the total impact of this attenuation in specific listening environments. DI provides a single number calculation that is representative of the frequency specific spatial attenuation properties that are displayed in directional patterns. The DI of hearing aids is of interest since it is assumed that it approximates the effective SNR for a condition in which the signal of interest originates directly in front of the hearing aid wearer and a fully diffuse noise field of the same total acoustic power is present. DI in most amplification systems designed for the hearing impaired

varies from approximately -3 dB to approximately +12 dB in some microphone array systems. Hearing aids that are equally sensitive to sound arriving from all angles (true omnidirectional) will have a DI = O dB - though when the hearing aid is used, the directivity pattern will never assume a perfect omnidirectional behavior. For sounds arriving directly in front of the listener, the DI will assume positive values and when sensitivity to sound is generally poorer for sounds arriving from directly in front of a listener, in comparison to sound arriving from all other angles, the DI will be negative. DI of hearing aids is usually calculated from two-dimensional directional patterns, three-dimensional directional patterns, or diffuse field versus free field measures. Beranek [1] proposed a method of calculating DI, usually used with three-dimensional directional pattern data.

In the clinical context, there is a more simple and faster method of evaluating and quantifying the directivity of a hearing aid. The front-to-back ratio (FBR) is the frequency-specific difference between the output level of a hearing aid in response to a sound source placed directly in front of a listener (0 degrees azimuth) versus that measured for the same sound source placed directly behind the listener (180 degrees azimuth). Once FBR measures are made for several hearing aids, clinic or patient/instrument specific normative values can be generated for comparison to future measurements. These data can be used to easily assess the functioning of the directional microphone in general, or the influence of patient specific factors such as venting.

On the behavioral aspect of the evaluation, the level of additional benefit directional hearing aids will provide to individuals having hearing loss is the most important to evaluate. It can be quantified using objective measures of speech recognition as well as subjective measures of the perception of sound quality, benefit, performance and satisfaction. By far the most common method for assessing the impact of hearing aids is the quantification of changes in speech recognition in noisy environments. When measured in traditional laboratory settings, directional benefit values ranging from approximately 5.5 dB to 11 dB and 40% to 70% have been reported in the literature. Studies, however, that have evaluated directional benefit in noisy environments designed to emulate difficult real-world conditions generally report values less than 6 dB and 40%. [39], [40], [42], [42]

2.2 Ambisonics

For the development of a virtual prototyping framework able to simulate a processed sound field with multi spatialized stimuli, and with independent control over their sound propagation, it is necessary a well discretized rendering of the spatial information around the listener's position. In this section, the encoding and binaural decoding is explained for the ambisonics format. For its optimization and efficiency [13], ambisonics is used for the reproduction of a full 3D virtual

acoustical space. As a sound reproduction technique, it involves a limited number of playback channels that, according to the holographic theory, can express the sound field as a superposition of plane waves. The mathematical formalist for the ambisonics system comes from the solution of the wave equation in a three-dimensional space.

The Kirchhoff-Helmholtz integral relates the pressure inside a source free volume of space to the pressure and velocity on the boundary at the surface. It is therefore possible to reproduce the original sound field by a determined number of loudspeaker signals. A variable geometry is render by designing a decoder that can weigh the sound pressure in the sphere to a finite number of loudspeakers array and also to binaural rendering over headphones. The minimum number of needed loudspeakers to represent the ambisonics order is given by the number of audio channels present in that order. Consequently, it can be shown that higher order ambisonics systems are increasingly accurate in the spatial information encoded.[36] The representation of the space is defined by a set of orthogonal basis functions and can be used to describe any function on the surface of a sphere, also named spherical harmonics.

2.2.1 Spherical Harmonics

The spherical harmonics represent the sound decomposition into frequency, radial and angular functions [51], into what leads to the Fourier-Bessel series [13].

$$p(\vec{r}) = \sum_{m=0}^{\infty} j^m j_m(kr) \sum_{0 \leq n \leq m, \sigma = \pm 1} B_{mn}^{\sigma} Y_{mn}^{\sigma}(\theta, \phi) \tag{2.2}$$

For each term of the order m, a radial, spherical Bessel function $j_m(kr)$ is associated with angular functions $Y_m^{\sigma}(\theta, \varphi)$ called spherical harmonics. [13], [51]

The aim is the re-synthesis of sound sources from particular spatial directions, either by reproducing dedicated ambisonics microphone recordings or synthetic signals. Within ambisonics, spherical coordinates are used, whereby φ is the azimuthal angle in mathematical positive orientation (counter-clockwise) and ϑ being the elevation angle with 0^o pointing to the equator and $+90^o$ pointing to the north pole. [14]

Considering an audio signal $f(t)$, which arrives from a certain direction $\theta = (\varphi, \vartheta)$, the representation of the surround audio signal $f(\varphi, \vartheta, t)$ is constructed using a spherical harmonic expansion up to a truncation order N.

$$f(\theta, \phi) = \sum_{n=0}^{N} \sum_{m=-n}^{n} Y_n^m(\theta, \phi) \phi_{nm}(t) \tag{2.3}$$

where Y_n^m represents the spherical harmonics of order n, degree m and $\phi_{nm}(t)$ the expansion coefficients. With increasing order N, the expansion results in a

more precise spatial representation. Spherical harmonics are composed of a normalization term $N_n^{|m|}$, the associated Legendre function P_n^m and the trigonometric function. [14], [51], [28]

$$Y_n^m(\theta,\phi) = N_n^{|m|} P_n^m(sin\vartheta) \begin{cases} sin\,|m|\,\varphi & m < 0 \\ cos\,|m|\,\varphi & m \geq 0 \end{cases} \tag{2.4}$$

$P_n^m(x)$ are the associated Legendre functions [53]. Considering the use of Resonance Audio as the software for spatial audio rendering, Ambisonic Channel Numbering (ACN) and SN3D normalization are used in equation 2.3. ACN defines the ordering sequence for the spherical harmonics channels as:

$$ACN = n^2 + n + m \tag{2.5}$$

And the normalization convention used is SN3D, often seen in combination with ACN, with the form:

$$N_{SN3D} = \sqrt{(2-\delta_m)}\frac{(n-|m|)!}{(n+|m|)!} \tag{2.6}$$

Using this index neatly defines a sequence for the spherical harmonics $Y_n^m(\theta,\phi) = Y_{ACN}(\theta,\phi)$ and the ambisonic signals $\phi_{ACN}(t)$ to stack them in a vector

$$y(\theta) = \begin{pmatrix} Y_0\theta \\ ... \\ Y_{N+1)^2-1}(\theta) \end{pmatrix} \tag{2.7}$$

The spherical domain components can be understood as the reconstruction of the wave field around the origin using a set number of microphones with multiple directivity patterns that define the magnitude of the signal and the direction of arrival. The higher the order of ambisonics, the more directivity patterns are assumed with a narrowed region of sensibility and thus a higher spatial resolution is rendered. [51], [13] (see figure 2.6)

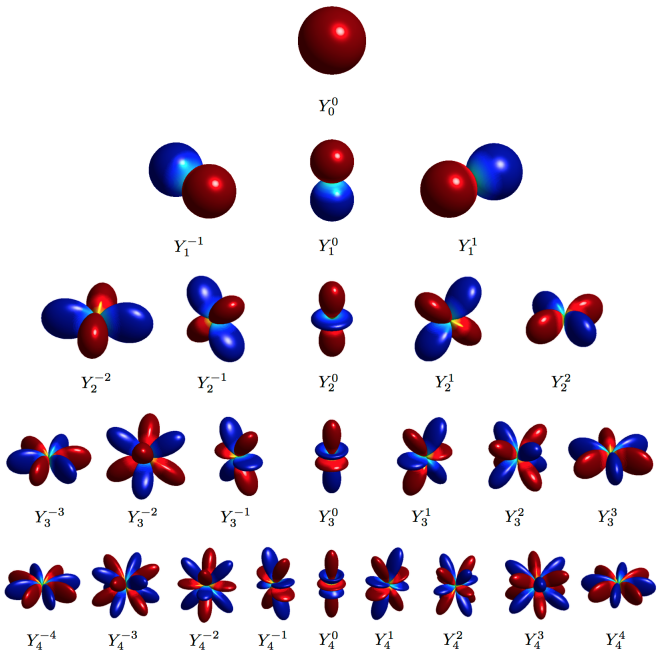


Figure 2.6: Directivity patterns of the real spherical harmonics up to order 4. Red color beams are positive amplitudes and blue colored represent the negative amplitudes

Higher Order Ambisonics (HOA) consider all the spherical domains above the truncation of N = 1. This representation requires $(N+1)^2$ spherical harmonics - HOA signals - and $(2N+1)$ channels for each ambisonics order.

2.2.2 Encoding of Virtual Sources

> Ambisonics directional encoding and decoding basically assumes that virtual sources as well as reproduction loudspeakers are in far field and radiate plane waves. [13]

Virtual sound sources can be reproduced in the HOA format using information regarding the source position and directivity. Daniel's HOA formulation [14] regards the spherical domain with real-valued spherical harmonics Y_n^m and $\mathbf{B}$ as the vector of real ambisonics signals B_n^m. The encoding can then be simplified to:

$$\mathbf{B} = S\mathbf{Y} \tag{2.8}$$

where S is the source signal, $\mathbf{Y}$ is the vector of real spherical harmonics Y_n^m and $\mathbf{B}$ the vector containing the real part of the ambisonics signals. B can be further described as a matrix of $(N+1)^2$ digital signals of length L. Introducing the component of distance ρ, the source signal can be encoded for near-field as:

$$B_n^m = SF_n(\rho, \omega)Y_n^m(\theta_S, \phi_S) \tag{2.9}$$

In this way, not only sources outside that are assumed to be far enough to have a contribution similar to a plane wave are encoded as well as sources that are enclosed by the rendered spherical domain, in the ear field, are encoded. This sources present a curvature in the wave front which allows the listener to perceive the source distance when moving in the sound field, independently from any room effect.

> Even for a still listener, the near field effect of close sources is perceptible through the emphasis of ILD (Interaural Level Difference).

2.2.3 Binaural decoding of HOA

Performing the encoding of the virtual audio sources in the ambisonics domain enables the representation of the full three-dimensional sound field, which can then be reproduced with a geometrical array of loudspeakers distributed on a sphere and radiating towards the origin of the sphere. [51] For the purpose of this experiment, and for the current market opportunities in the field of virtual reality, augmented reality and mixed reality, the binauralization of the sound field is in fact the most practical choice of reproduction. Two different approaches can be used for the binaural decoding [51] over headphones: set an array of virtual loudspeakers that would form the spherical reproduction as if it was an array of real loudspeakers, and assign two head-related transfer functions (HRTF) to each loudspeakers, for each ear. The output signal for each ear will then be the sum of L loudspeaker signals ($\sum_{n=0}^{N}\sum_{m=-n}^{n} B_n^m(\omega)D_{n,l}^m$ convolved with the corresponding HRTFs, $H_{l,left}(\omega)$,$H_{l,right}(\omega)$:

$$S_{ear}(\omega) = \sum_{l=1}^{L} H_l(\omega)(\sum_{n=0}^{N}\sum_{m=-n}^{n} B_n^m(\omega)D_{n,l}^m) \tag{2.10}$$

The second approach to binaural HOA reproduction consists on the pre-computation of the spherical harmonics-based HRTFs $H_n^m(\omega)$ by solving the equation:

$$H(\theta, \phi, \omega) = \sum_{n=0}^{\infty}\sum_{m=-n}^{n} H_n^m(\omega)Y_n^m(\theta, \phi) \tag{2.11}$$

Where the spherical harmonics coefficients can be determined by direct integration or using the least square solution. The decoding process is defined by the loudspeaker arrangement only. Therefore, to avoid ill conditioning or even singularities in the decoder matrix, it is important to distribute the loudspeakers as uniformly as possible over the spheres surface. [36]

2.3 Case study: Resonance Audio

In this project, ambisonics was used for the rending of complete virtual sound sources with specific spatial information. *Resonance Audio* (RA) [1] performed the encoding of the virtual audio sources, tracking their spatial position in terms of direction and distance (r, θ, ϕ) and encoding it directly with the spherical harmonics of third order.

RA uses ACN channel ordering in combination with SN3D normalization. The encoding in the spherical domain is built using the second approach described in the previous section, where the HRTF are pre-computed to the spherical domain, allowing to project multiple sound objects in the ambisonics soundfield. In the case of the third order ambisonics, a Lebedev grid was used to distribute the matrix of virtual loudspeakers in the most uniform way. From this grid, a matrix of seventeen angles was used in Matlab[2] to pre-compute the spherical harmonics of the third order of HOA [24] [30]. The RA package comes with a Unity API built on top of the Unity audio engine which enables the use of HOA encoding up to third order. Bellow is a diagram of the classes defined by RA Unity API that were used for this project.

Classes	
ResonanceAudio	This is the main Resonance Audio class that communicates with the native code implementation of the audio system.
ResonanceAudioListener	Resonance Audio listener component that enhances AudioListener to provide advanced spatial audio features.

Table 2.1: Resonance Audio classes used in the project from the Resonance Audio API for Unity

For the purpose of mapping the interaction, not all the the properties of the each class were used. Resonance Audio Class was used as a connection between the functions of Resonance Audio Listener and Resonance Audio Source and the respective native components in Unity as well as for the GUI rendering of the gizmos that represent the polar patterns of both elements in the Unity Scene (see figure 2.7). The attributes of limit distance, and the range of the gain are also computed through this function.

The Resonance Audio Listener class gives access to extra features of control from inside the Audio Listener native class from Unity. The attributes used were the global sound pressure level (in dB) and the listener directivity parameters alpha - controlling the polar pattern of the listening sensibility - and sharpness - to adjust the width of the beam defined by the polar pattern of alpha.

An important note to remember is that the Audio Listener class works independently for each audio source. In Unity, the relation between audio listener and

[1] https://developers.google.com/resonance-audio/
[2] https://www.mathworks.com/products/matlab.html

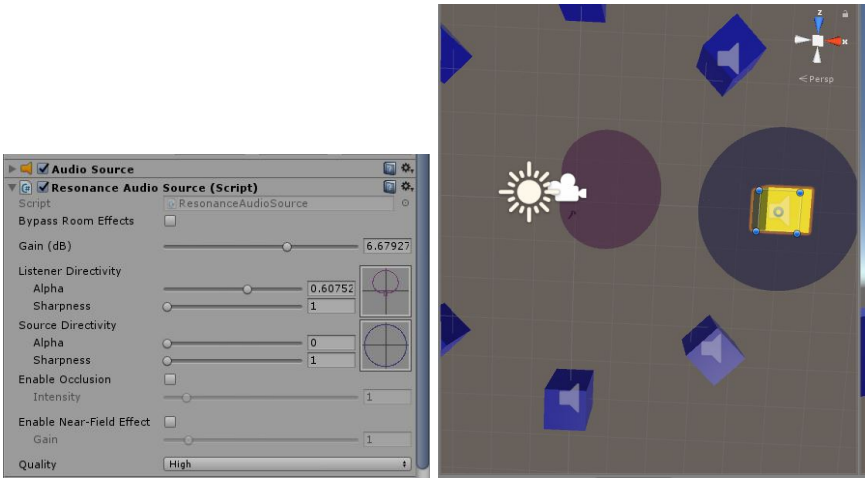


Figure 2.7: Resonance Audio Listener and Resonance Audio Source gizmos representation. On the left side, is the GUI for the RA plug-in for Unity, where both audio listener's gizmo for directivity and the audio source's gizmo can be controlled. On the right side, the scene view where it's possible to see the same patterns and their representation in the space.

audio source is a closed relation and source specific, meaning that any change in the parameters above described will only affect the source for which they were changed.

Chapter 3

Experiment

The initial proposal of this experiment was investigate if the interaction models implemented could improve the ability of auditory attention and to understand if the participants would be able to perform better during the test, both in terms of task results as well as in terms of quality of experience (user experience).

Virtual reality provided an optimal platform for developing the dynamic interactions for head and eye movements as well as the hand remote controller, and to collect interaction-specific information of each of them. The virtual generated environment allowed to separate these interactions or at least, to register independent information for each of them and understand possible behavioral patterns in the way the task was solved as well as correlation between motion and performance of the users. The possibility to isolate these modalities of listening and explore how participants performed with beamforming control the interactions offered was a very insightful and interesting asset of VR.

The following sections describe the technical implementation of the virtual scenarios created and the mapping control of the interactions, the experiment design and protocol.

3.1 Technical implementations

The test was built for a VR setup using Unity, a game engine software that allowed the interactive structure required for the different stages of the experiment as well as for the use of Resonance Audio API. Two different sets were created: one with virtual cubes and another one where 360 videos were loaded to the scene, simulating a multi-talker environment. To track the participant during the performance, both head and controller coordinates for rotation and translation as well as eye gaze were captured every 3 frames using OSC communication protocol with Max/MSP - to store the values in a text file. Beyond these values, also duration of the test, sequence used during each interaction test, source selected for the task

and calibrated parameters' values (for directivity and sound pressure level) were also tracked with.

The experiment ran using HTC Vive[1] headset with a tracked area of approximately $3m^2$ [2], allowing the participant to have a slight degree of motion freedom in space. For the binaural audio reproduction, a pair of Sennheiser HD600 was used combined with a headphone equalization filter, *Equalizer APO* [3] that introduced a compensation to filter out the influence of the pinnae information that is encoded in the HRTF databased of the Resonance Audio plug-in. This way for the spatial audio rendering only the influence on the frequency response from the participant's outer ear was taken into consideration.

3.1.1 360 Video Recordings

For the 360 video recording setup, 8 speakers were placed around the listener's position in an anechoic room. The recording contained a 360 video, captured with *Garmin Virb 360*[4] camera, of the speakers reading the sentences and the audio signal from each speaker taken with clip microphones, DPA SC4060 [5]. (see figure 3.2) The signal cross-feed between microphones was around 15 dB and was later reduced with equalization and compression to close to 10 dB difference between main signal and "bleeding" signal. After the post-editing of the videos and the audios, the output signal was measured to check if it complied with a sound pressure level of 60 dB SPL, equivalent to a conversational signal level.

Figure 3.1: Recording position of the speakers

These 8 speakers read sets of sentences available from IEEE Recommended Practice for Speech Quality Measurements appendix list [19], also known as the famous Harvard Sentences, used in many different fields of audio engineering (e.g.

[1]https://www.vive.com/us/product/vive-virtual-reality-system/
[2]https://www.digitaltrends.com/virtual-reality/oculus-rift-vs-htc-vive/
[3]https://sourceforge.net/projects/equalizerapo/
[4]https://buy.garmin.com/da-DK/DK/p/562010
[5]https://www.dpamicrophones.com/dscreet/4060-series-miniature-omnidirectional-microphone

speech-to-text software, for cochlear implants). These sentences are considered standard and as a optimal research material for the fact that all the word lists that were phonetically balanced, meaning that the frequency of sounds in these lists matched that of natural language. If you used these words, you were sure to hit all the noises a person would typically hear in a conversation. The sentences are deliberately simple and short-monosyllabic words punctuated by exactly one two syllable word sentence.

There are four female and four male speakers, paired in groups of two, giving a total of four pairs per recording. The sentences in each set (called sequence) were randomized seven times. Not just the order of the sentences is randomized for each sequence, but also the sets are attributed randomly for each speaker in the circle during the recording – all the randomizations were made using latin squares to achieve the most balanced distribution between subjects. (refer to the Appendix to access to the experiment material - sequences, ordering and interaction orders, and test instructions)

3.1.2 Interaction metaphors and mapping control

The three interactions were developed to work in similar manner. The main process in the interaction pipeline regards the way the participant focus its attention to a certain speaker/source as its focus source which is intimately related to the limits of the user's movement. The three dynamic parameters that can be changed by the user – directivity alpha, sharpness and gain level – can reinforce the natural/intuitive use of each of the interactions. During the training period these values could be changed by using the hand controller and a specific mapping adapted:

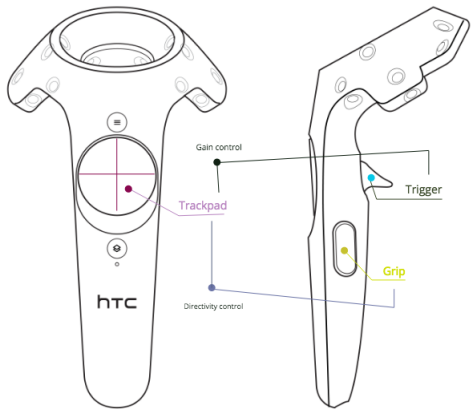


Figure 3.2: HTC Vive controller and mapping combinations for parameters's control.

By pressing different button combinations in the hand controller, the user was able to adjust directivity parameters and gain parameters independently. The grip and up or down in the trackpad would change the directivity pattern, and left and right would affect the width of the pattern. On the other hand, trigger and up or down on the trackpad with increase or decrease the gain value of the associated source.

The directivity alpha parameter can be adjust from a full omnidirectional pattern ($\alpha = 0$) to a maximum of a bipolar pattern ($\alpha = 1$), being $\alpha = 0.5$ equivalent to a perfect cardioid sensibility pattern. The idea behind this directivity pattern coefficient is to shape the directivity of the user to each source of signal independently, as if the user was wearing a pair of microphones on the ears. The equation for the calculation of this polar patterns is:

$$P(\theta) = |(1 - \alpha + \alpha.cos(\theta))|^{sharpness} \tag{3.1}$$

where α can assume a value in the range of 0 to 1 and sharpness regards the width of the beam, being a value of 1 equivalent to the natural sensibility and 10 the maximum value of narrow-th possible.

A smooth function was introduced for when the user changes focus source during the testing period, enabling a progressive transition between the new selected source and the previous one with approximately half a second. Some of the feedback regarding this implementation complain about the duration of this cross-fading and implied that a faster transition could have been more natural/efficient or even the presence of no smoothing function. However, studies such as the one by Hamacher [18] suggest the fading from one directivity pattern to another and that the use of a simple step function, or a sudden "off/on" switch of the signal processing that affect the directionality of the microphones are considered irritating and unpleasant.

The three interaction can be organized as natural/embodied interactions (head rotation and eye tracking), and artificial/reinforcing interaction (hand controller).

Head Movement

When using the head interaction the user is able to choose/focus on a specific target by rotating and position the head directly in front of the area of interest. The listener's area is divided into as many areas of interest as the number of stimuli that should be tracked. In the case of this experiment, since there were 8 speakers/sources, the listener's area was divided into eight equally spaced slices. To avoid "selection noise" when the participant moves the head, the interaction waits one second to understand what is the source in front of the listener. This way, the listener is able to freely move the head without worrying to be constantly selecting random focus sources.

Eye Movement

The eye interaction is related to the position of the eyes in a two-dimension plane (see figure 3.3 and the intersection of the eye gaze with position of the sources in the space. The implementation in Unity was possible using the eye-tracking system by Pupil Labs [6] and its Unity library. The interaction is defined by creating a collision vector between the eye gazing point and the possible collision with an object in the virtual environment. There were other prominent implementations that could define a heat map and see the difference in the coloration between background and speaker/virtual cube. However, the current implementation was enough to provide a simple and efficient interaction. The listener was able to select an audio source by looking directly into it. Contrary to the head movement, this interaction enable the participant to choose between one out of two sources in the same field of view. Nevertheless, the eye movement is still dependent on the head rotation to be able to select sources from behind.

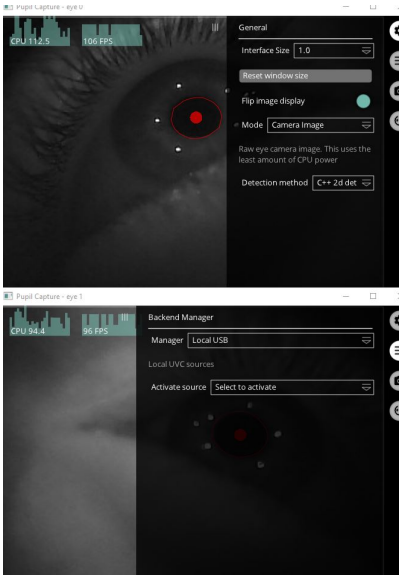


Figure 3.3: Eye-tracking screen-shot from the Pupil Labs application.

Hand controller

The hand controller was used as a pointer interaction where the participant would point the controller towards the source of interest. Just like the head, the controller is tracked in space and its relative position to the meaningful sources around the listener is computed so that the area the controller occupies, always enables the focusing of a certain source. Contrary to the head interaction however, the controller

[6] https://pupil-labs.com

has an instantaneous selection timing, so the user can very fast change between focus sources. It seemed natural during the implementation of this interaction that since the controller is considered as an extension of the body movement, it isn't directly affect by the limits of the body rotation, and it's in fact this separation that enables it to be consider as an extension. The participant could be facing a particular source while pointing the controller to another, giving the system the information that the other source would be more relevant than the one right in front.

3.2 Experiment Protocol

The experiment can be described as a task-based testing, where the subject is asked to find pairs in a group of eight speakers. To complete this task, she/he is able to dynamically change the shape and the width of a directivity beam as well as the sound pressure level for each of the voices displaced in the space. These eight sources/speakers are distributed in a perfect circle, separated by a 45° angular distance. (see fig. 3.4)

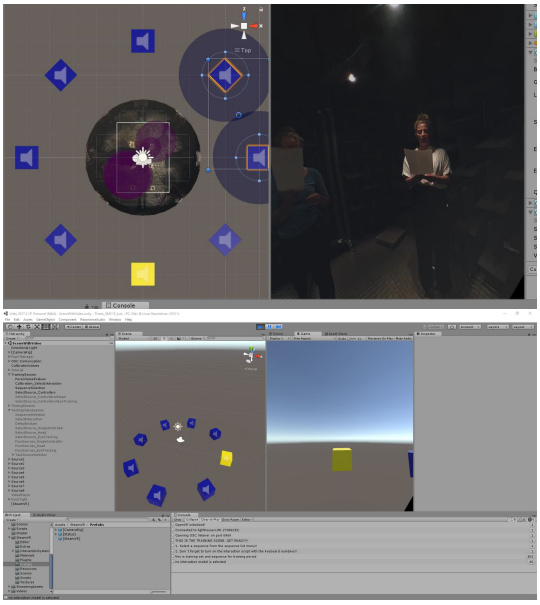


Figure 3.4: Speakers' displacement. On the left the virtual rendered cubes and on the right, a 360 video recording

For the purpose of the experiment, three different types of interaction were considered: head rotation, hand controller and eye tracking were individually tested during the experiment for each participant. These three interaction models were also randomized, so that no consideration was given to a particular order during the tests. The subjects were always informed of which interaction was to used, but

no further indications regarding position of the pairs, nor which pairs were correct was given.

3.2.1 First and Second Experiments

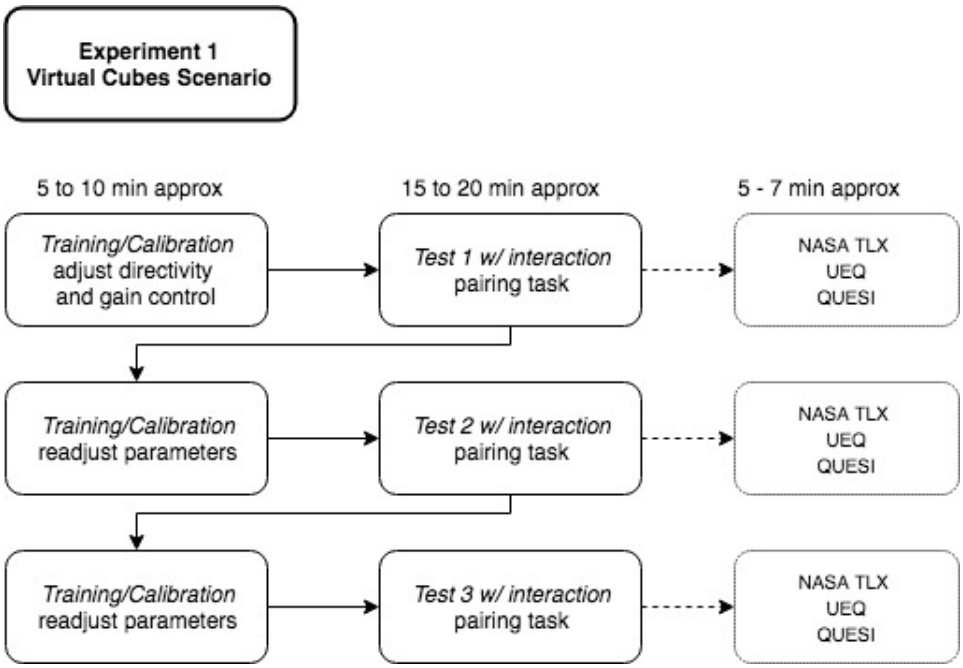


Figure 3.5: First experiment scenario schematics - virtual cubes

The experiment procedure was broken down into two different test blocks (see fig. 3.5 and fig. 3.6). A first test where the users were required to perform a small training and calibration of the sound field around them, followed by a test with virtual cubes playing the sets of sentences previously described. In this stage, the goal was to offer the participants a limited biased environment where the main cues were restrictedly auditory. In this sense, the performance of the tasks would be above all a test to the participant's ability to understand the messages and cognitive strategies. The second experiment was built so that the same participants performed the same pairing task and listening to the last sentence task but in a 360 video environment where they could see real people saying the sentences. This second experiment was designed to understand if the virtual prototype enhanced the listening experience in a real world environment (with the 360 video) by using the same values as the ones defined by the participant in the first experiment. Could the user performer better results with the presence of a visual cue? Would the values previously defined still be good in this scenario or modifications would be advised?

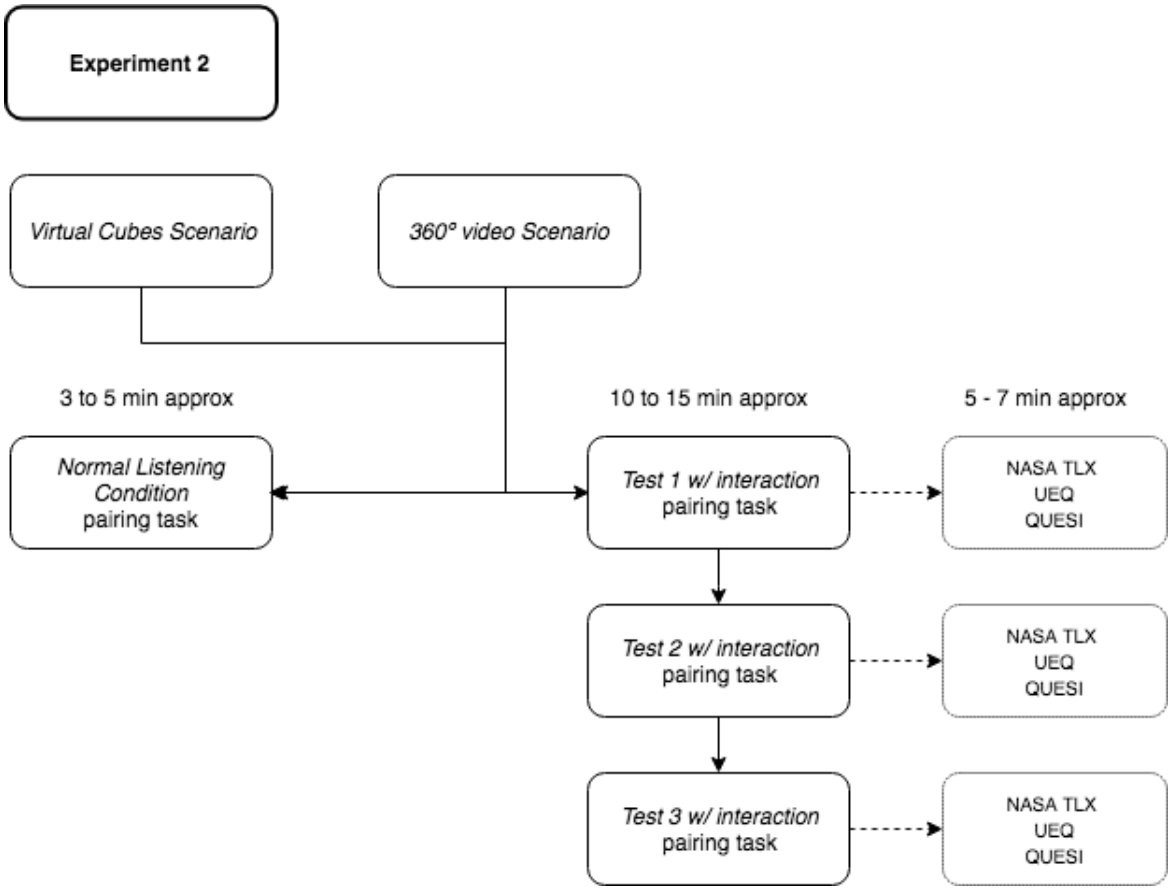


Figure 3.6: Second experiment scenario schematics - 360 videos

Both tests also included a normal listening condition, where the users were given no interaction tools and asked to perform the tasks using only their normal listening abilities.

Before performing the task, the subjects had to undergo a training period in a scene consisting only of virtual cubes displayed around the listener. (see figure 3.7) The goal of the training period was for them to evaluate what could be the optimal values of directivity parameters for a natural intelligibility. In that sense, they were asked to change the values of directivity alpha and sharpness of the listener directivity and the gain level of the stimuli, independently, so that they could clearly understand what was being said by the voice in one of the cubes (the one in yellow on figure 3.7) and perceive its spatial position when rotating in the scene. These values were then used as the default values for the single interaction metaphors of head rotation, eye tracking and hand controller, which relied on fixed values. To do this, the participants would require to point the hand controller (which wouldn't work as a interaction tool for beamforming control but

as a remote control that they could use to change the parameters of control. The range of gain control was defined to be a continuous value between -20 dB FS and 20 dB FS, while the directivity was enclosed in a limit of 0 to 1 for the alpha component (described above) and from 1 to 10 for the sharpness. To reduce the time spent on understanding the different systems and the different mappings, the participants were asked to do a short tutorial where they could get familiar with how the controls work and what was the audible effect of changing both directivity and gain values for three different sources in the space.

After the training period, participants would start the test. For each interaction, there were 2 tasks: pairing the sources according to which set they belong to and repeat the last sentence from one of the sets. The playback of the sequences and the 360 videos ran in a loop until the subject was able to give an answer to the tasks. After performing the tasks for each interaction, the participants were asked to answer three questionnaires: a psychological evaluation of the metal effort to solve the task (NASA TLX) [17], and two user experiment tests User Experience Questionnaire (UEQ) [44] and Questionnaire for Measuring the Subjective Consequences of Intuitive Use (QUESI) [35] - analysis of these questionnaires can be found in the next section.

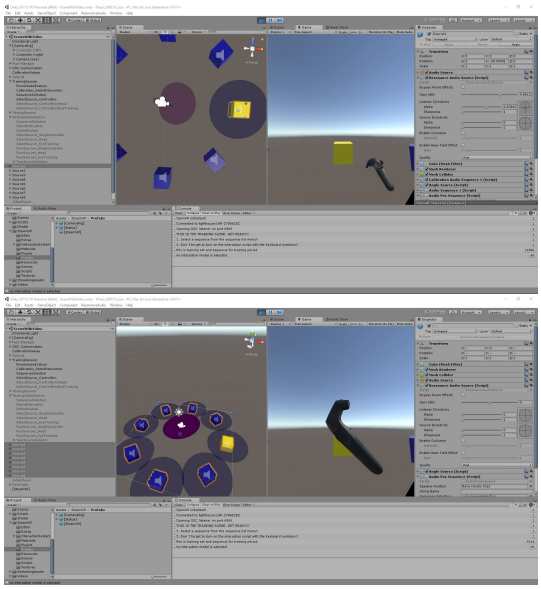


Figure 3.7: Training Setup

To reduce the time spent on understanding the different systems, the participants were asked to do a brief tutorial where they could get familiar with how the controls. (see figure 3.8). This tutorial was divided into four levels: an introductory level to the hand controller, where they were asked to increase the volume of a cube and shape the directivity of another cube into a very narrow bipolar pattern.

On the next three levels, the cubes were already calibrated with fixed values for both directivity and sound pressure level so that the subject would only need to try to move around using the sonic interaction control to select a specific focus source.

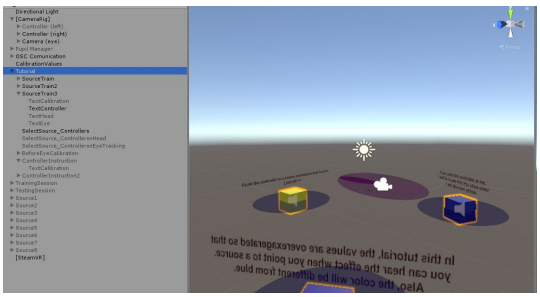


Figure 3.8: Tutorial scene, before training stage

Chapter 4

Analysis and Results

This chapter looks into the data collected, what the different variables might represent and how it was framed to help solve the initial problem statement. On the next chapter, a deeper interpretation of the results is presented.

During both experiments, participants were asked to answer three different questionnaires after performing the task for each interaction metaphor. These three questionnaires were used to assess two user qualities of the experiments:

- Measuring the product experience with UEQ in terms of the rating of attractiveness, goal driven pragmatic aspects (perspicuity, efficiency and dependability) and hedonic attributes, related with stimulation and novelty of the experience. [44]
- Measurement of the intuitive use or the subconscious application of prior knowledge that leads to effective interaction, using the QUESI form, a set of 14 questions were asked regarding the use of the different systems.
- Quantifying the mental effort regarding the task workload based on six factors: mental, physical and temporal demand, effort, performance and frustration [17], suggested with the NASA Task Load Index form.

Beyond the information collected with these assessment tests, real-time data from Unity play-mode was also gathered using OSC protocol and Max/MSP. This data stream was composed of multiple variables and it's detailed in the table 4.1. It was mainly used to keep track of the performance of the participants in terms of task results: how many pairs of voices they were able to find and group together, what values did they opt for during calibration for both directivity parameters and sound pressure level as well as how long did it take to finish the task.

OSC Communication Unity - Max MSP	
Headset	Coordinate values of the headset in space and rotation in relation to the center axis of the Unity environment
Hand controller	Coordinate values of the hand controller in space and also in terms of rotation
Eye tracker	two-coordinate system (XY) for the position of the eye gaze in a parallel plane to the subject's eyes
Session	If it's training, testing with virtual cubes or with 360 videos
Sequence	sequence number for the pairs. According to each sequence, there will be different order of the pairs and the voices that are paired
Interaction	Interaction used for testing
Source Selected	What source was selected
Directivity Alpha	Value of the directivity alpha for each of the 8 sources in the scene
Directivity Sharpness	Value of the directivity sharpness for each of the 8 sources in the scene
Gain Value	Value of the gain for each of the 8 sources in the scene
Paired Sources	Sequence of pairs found by the participant in the test
Duration of the Task	Timer with the time stamps for when each pair is found and the total duration of the test

Table 4.1: Data stream collected using OSC communication to Max/MSP

4.0.1 Sample groups

For the two experiments, a total number of seventeen participants were tested. From these seventeen participants, seven of them also performer the second experiment with 360 video scenarios, using the values they previously calibrated in the first experiment. The average age of the sample was 27 years-old - with a standard deviation of 6.2 years - and five out of the seventeen participants were women. From these, in the second experiment, two were woman and five participants were men. Regarding the exposure to virtual reality, 50% of the participants stated to already have tried interactive games and demos, 45% had tried 360 videos. None of the participants self-reported any hearing deficit or impairment.

4.0.2 Real-time collected data analysis

From the previous described real-time data stream, in table 4.1, one of the most interesting analysis to make regards participants' precision to complete the task and evaluate the frequency and rank of their scores, which is equivalent to the number of pairs of voices found in the scene. In this perspective, it is important to understand if it's possible to see any correlation between participants that scored higher with a specific interaction condition, or with specific calibration values or even if the visual cue of the 360 videos could enhance in any way their performances.

Before deciding which statistical evaluation to use, some hypothesis were formulated.

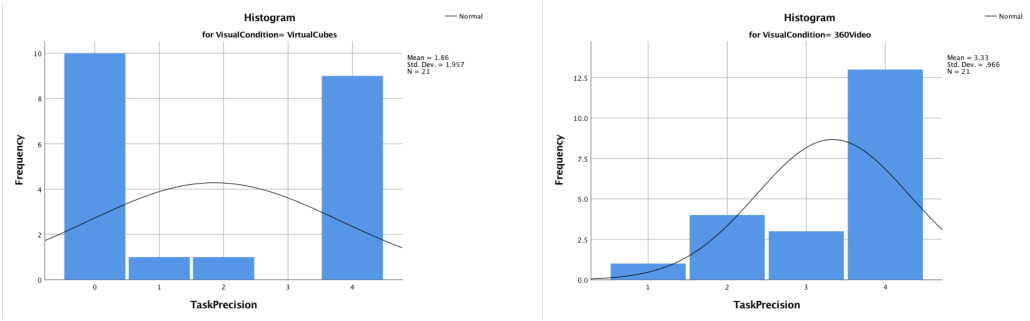
- **Visual Feedback Precision**: comparison of task precision - the number of pairs the participants were able to find - between the first experiment with

the virtual cubes scenario and the second experiment with the 360 video scenario.

- **Task Precision Between Interactions for Experiment 1**: comparison of the task precision between interaction conditions in the first experiment
- **Task Precision Between Interactions for Experiment 2**: comparison of the task precision between interaction conditions in the second experiment

The null hypothesis for all the above can be consider the same, stating that there are no significant differences between conditions for the dependent variable in analysis. This can be rephrased as no relevant difference was found between interaction models or even between scenarios.

To decide on what evaluation to use it was necessary to observe if the parametric assumptions were satisfied or if there was no coefficient of normality in the data collected. The histograms in figure 4.1 prove indeed that this assumptions are not guaranteed and thus non-parametric tests were performed (Mann-Whitney Test for the two conditions of Visual Biased Precision and Kruskal-Wallis H test for the three conditions - the comparison between the three interaction models). It is also worth noting that the ratios for the descriptives of kurtosis and skewness also provided values outside the range of [-1.96; 1.96].



(a) Distribution of scores on virtual rendered cubes scenarios

(b) Distribution of scores on 360° video scenarios

Figure 4.1: Histograms of the distribution of participants' scores over scenarios from the two experiments.

Two out of the three hypothesis proven to be statistically significant. For the Visual Biased Precision comparison, the null hypothesis was rejected proving that participants had higher scores during the second experiment, over 360 video scenarios, with a significance level of 0.018, lower than the asymptotic significance of 0.05. The precision rate between interactions for the first experiment (with virtual cubes scenario) also proven to reject the null hypothesis showing a significance

level of 0.032, with particular high scores for the hand controller interaction, compared to the eye tracking and head rotation that followed with lower rank means.

When comparing the three interactions in the second experiment, no statistical significance was considered though the head rotation interaction assumes to have a better performance than the eye tracking. In this scenario, the head rotation interaction is in fact the one which out-performs the others interactions. (refer to figure 4.2)

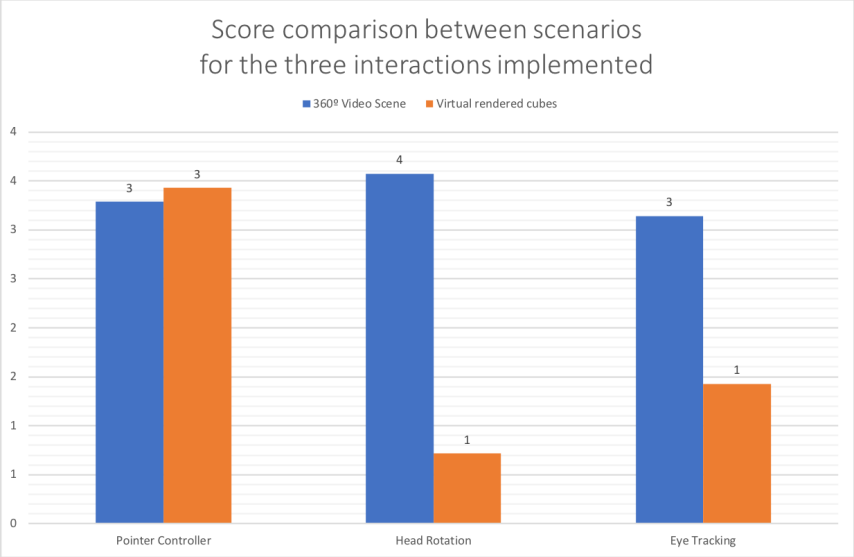


Figure 4.2: Histogram of the precision of participants between interactions, for both scenarios. The orange bars represent the second experiment with 360° videos and the blue bars represent the precision on the first experiment with the virtual rendered cubes.

A further analysis into the first experiment data (where the differences between interactions were meaningful) and picking only the best and the worst task scores, it is possible to study if any plausible correlations can be found between an average of the calibration values during training for each of these performances. Using Pearson's correlation analysis, it is evident that the best performances are indeed extremely correlated with the average values of calibration, with a positive 0.95 Pearson correlation value, for a significance level lower than 0.01. On the contrary, the worst scores proven to not be statistically significant, which to a certain degree can be understandable as there are multiple possibilities to achieve low scores. (see figure 4.3)

4.0.3 UEQ and QUESI-A

The User Experience Questionnaire and the Questionnaire for Measuring the Subjective Consequences of Intuitive Use are used as a measure of statistical com-

Correlations

		Average	Subject3	Subject9	Subject2	Subject10
Average	Pearson Correlation	1	.976**	.958**	-.304	.137
	Sig. (2-tailed)		.000	.000	.464	.747
	N	8	8	8	8	8
Subject3 (best score)	Pearson Correlation	.976**	1	.951**	-.345	.041
	Sig. (2-tailed)	.000		.000	.403	.922
	N	8	8	8	8	8
Subject9 (best score)	Pearson Correlation	.958**	.951**	1	-.496	.057
	Sig. (2-tailed)	.000	.000		.212	.893
	N	8	8	8	8	8
Subject2 (worst score)	Pearson Correlation	-.304	-.345	-.496	1	-.391
	Sig. (2-tailed)	.464	.403	.212		.338
	N	8	8	8	8	8
Subject10 (worst score)	Pearson Correlation	.137	.041	.057	-.391	1
	Sig. (2-tailed)	.747	.922	.893	.338	
	N	8	8	8	8	8

**. Correlation is significant at the 0.01 level (2-tailed).

Figure 4.3: Pearson's correlation table. The two best and the two worst scores were correlated with the average calibration values

parison between different products. For this research, it was also important to understand if any meaningful information could be retrieve from the participants experience when trying to solve the task, if there was any general tendency to appreciate the use of one type of interaction over another, if one shown to have more potential or more natural/intuitive mapping.

In UEQ, the items have the form of a semantic differential, being represented by two terms with opposite meanings. For the experiment the terms used were:

UEQ Evaluation	
Obstructive	Supportive
Complicated	Easy
Inefficient	Efficient
Confusing	Clear
Boring	Exiting
not interesting	Interesting
Conventional	Inventive
Usual	Leading edge

Table 4.2: Semantical values attributed

Again, in this situation, not all the parametric assumptions were satisfied and so a Kruskal-Wallis H test shows that the UEQ test has no statistical significance (figure 4.5), and the null hypothesis that there is no difference was not rejected.

However, looking at the mean values, in figure 4.4, some observations in the treatment of the data for each interaction can be done:

- The scores for all the attributes are always above 3 points, with a global mean of 4.
- Eye interaction is considered exciting, interesting, inventive and leading edge in comparison to the hand controller.
- The head rotation is also considered interesting

Controller	4	4	4	4	3	3	3	3
Head	4	4	4	4	3	4	3	3
Eye	4	4	4	4	4	4	4	4
Interaction with highest score	All	All	All	All	3	2,3	3	3

Figure 4.4: The table shows the average ratings of each question for the three interactions for the UEQ. In the yellow row, interactions are numbered from 1 - controller to 3 - eye tracking interaction

Test Statistics [a,b]

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
Kruskal-Wallis H	.024	.000	.093	.684	1.931	.070	3.981	2.421
df	2	2	2	2	2	2	2	2
Asymp. Sig.	.988	1.000	.955	.710	.381	.966	.137	.298

a. Kruskal Wallis Test
b. Grouping Variable: Interaction

Figure 4.5: The Kruskal-Wallis H test shows no statistical significance for the any of the questions of the UEQ for any interaction model.

The QUESI evaluation relates to the intuitive use of a system [35] trying to quantify subjective consequences. In a sense, it can be a seen as a combination of both the UEQ and the NASA TLX as an attempt to reinforce the analysis of both these tests. According to the authors [35], there are five major intuitive criteria of possible evaluation (refer to table 4.3).

Again, for this questionnaire, not significant level was found, which means that no difference in how intuitive the user found the use of the three interaction models. Again, it is possible to understand that the hand controller interaction shown to be the easiest interaction to perform the task, where the users shown little effort trying to understand how the system worked. (see figure 4.6)

QUESI
Low subjective mental workload
High perceived achievement of goals
Low perceived effort of learning
High familiarity
Low perceived error rate

Table 4.3: Intuitive use criteria

Figure 4.6: The table shows the interaction with highest scores for each question of the QUESI form.

4.0.4 NASA Task Load Index (TLX)

The NASA Task Load Index test was introduced to measure the possible workload on the participant while performing the pairing task during the experiments. The participant uses numerical ratings for each scale that reflect the magnitude of that factor in a given task. [17] These factors are used to analyze what were the most important contributors of the workload. They can be grouped as seen in the table 4.4).

NASA Task Load Index		
Mental Demand	Low/High	How much mental and perceptual activity, was required (thinking. deciding, calculating, remembering. looking, searching, etc)
Physical Demand	Low/High	How much physical activity was required (pushing, pulling, turning. controlling, activating, etc.)?
Temporal Demand	Low/High	How much time pressure did you feel, due to the rate or pace at which the, tasks or task elements occurred?
Performance	Low/High	How successful do you think you were in accomplishing the goals of the task, set by the experimenter (or yourself)?
Effort	Low/High	How hard did you have to work (mentally and physically) to accomplish your level of performance?
Frustration Level	Low/High	How insecure, discouraged, irritated. stressed and annoyed versus secure,,gratified, content, relaxed and complacent did you feel during the task?

Table 4.4: The six major factors defined by the NASA TLX test. They are rateed from low to high in a 5-point likert scale.

From the analysis of the collected data from both experiments, a Kruskal-Wallis H test was again performed for each factor, with no significant difference found between interactions (see figure 4.7). It is possible to observe however that the mean ranks (see figure 4.8) of the hand controller interaction proven to be the best in terms of the performance factor, providing some insight that perhaps participants were more confident while using this interaction to solve the pairing task.

On the other hand, the eye tracking interaction and the head rotation interaction shown to be more physically and mentally demanding. Perhaps further testing could provide more information if performance confidence is associated with the effort required to perform a task.

Test Statistics [a,b]

	TLX1	TLX2	TLX3	TLX4	TLX5	TLX6
Kruskal-Wallis H	1.058	1.705	.683	3.387	1.162	.813
df	2	2	2	2	2	2
Asymp. Sig.	.589	.426	.711	.184	.559	.666

a. Kruskal Wallis Test
b. Grouping Variable: Interaction

Figure 4.7: The table shows the average ratings of each question for the three interactions for the UEQ. In the yellow row, interactions are numbered from 1 - controller to 3 - eye tracking interaction

4.0.5 Spatial data collected from questionnaire and feedback

Beyond the data collected in real-time and the questionnaires above described, the participants were also asked two extra questions regarding their cognitive strategies to perform the task: "what type of voice was more easy to follow/understand?" and "what area of the sound field did they focus more their attention?". 87% of the participants replied to be more sensitive to the frontal area (perhaps related to their field of vision and as a initial stage of recognition of the spatial context) and 37% also felt more comfortable listening to the male voice in opposition to 25% who found more easy to follow the female voice.

NASA TLX : Mean Ranks		
	Interaction	Mean Rank
TLX1	controller	35.40
	head	39.85
	eyeTracking	34.25
	Total	
TLX2	controller	32.13
	head	38.31
	eyeTracking	39.06
	Total	
TLX3	controller	34.85
	head	35.44
	eyeTracking	39.21
	Total	
TLX4	controller	41.19
	head	37.75
	eyeTracking	30.56
	Total	
TLX5	controller	34.42
	head	35.04
	eyeTracking	40.04
	Total	
TLX6	controller	33.58
	head	38.63
	eyeTracking	37.29
	Total	

Figure 4.8: NASA TLX mean rank table for each factor for all the three interaction models.

Chapter 5

Discussion

The listening experience is modulated by a variable number of cognitive, bodily and external conditions that affect both the physical characteristics of the incoming signals as well as the listener's perception and intelligibility of the message. In situations such as the cocktail party scenarios, these conditions might have a strong impact in communication specially in the case of hearing impaired people. For this reason, it becomes important to understand to what degree could these conditions be controlled and and what tools could there be to enhance the listening processes that humans already have.

To perform this study, a virtual prototyping was developed using virtual reality rendering. This prototype would enable the participants to use different interaction metaphors and analyze real-time data from their experiments.

In this chapter, two different views of the project are formulated, questioning the choices made, the results granted and what could be the next steps for this kind of research.

5.1 Towards a super hearing experience: part I

The protocol developed for this study is presented in a two-part experiment. In both scenarios (both with virtual cubes and with 360 videos), participants were asked to perform the same task: find the four pairs of sources that are saying exactly the same sentences. To accomplish this task, they were given the chance to dynamically control the values of directivity pattern and sound pressure level for each source, during the training period. In this sense, different possible combinations were made by each participant, and in fact, after each test, participants were able, if they intended to do so, to change the values again, readjusting their calibration after trying it in the test. It is possible to understand now, that these three degrees of freedom combined with the three interaction conditions and repeated over two different scenarios transforms the experiment into a mixed-condition test.

The data collected and the statistical results cannot be defined as a consequence of the participant's exposure to one single condition but in fact to the combination of multiple factors that can influence the listener's perception. Another extra factor that affected the tests was described by some of the participants who observed that the test not only required them to be very focus on what they were trying to listen but as well as to discover the best strategy to find the pairs - a cognitive strategy to remember "who said what", bringing other levels of cognition to the experiment, than the ones purely related with the listening processing and speech recognition (see chapter Fundamentals for further reading on this subject).

Nevertheless, the study can be considered as an initial virtual prototype, layering the foundation for further research in this path. Giving that external conditions are considered treated before hand, the focus remains only on how the subject of the test uses the interaction tools provided to enhance its natural abilities for auditory attention. It can be assumed that to a certain extent this test provides a solid ground for the ecological validity of the experiment, but lacks on providing real direct causes of action over the choices in the participants' tasks.

Assuming this context, there are still some patterns that provide clues to what is the optimal way to present single audio streams to the listener. Looking at the first experiment, the statistical analysis shows a significant better performance of the hand controller over the other interaction metaphors, with an average of 3 pairs found, in comparison to 2 pairs for eye tracking and only 1 for the with the head rotation. This result stands out for the reason that it was in fact the less embodied interaction that provided the most efficient performance. Some of the feedback collected indicates that participants saw this interaction more as game-like, where perhaps in real-life it wouldn't provide the same level of comfort and in fact could be troublesome to constantly be pointing at someone to follow their voices, in VR it was actually simple and easy to use and also provided the ability to have a certain degree of freedom from the beam sensitivity of the pointer and face a different source that they wish to listen simultaneously. The head rotation and the eye movement could have been degraded in their scores for the fact that visually there was no relevant cues to help guiding the participants (they were only seeing eight cubes displayed in a circle), which saw much better performances during the second experiment - where participants confirm to be lip reading, and in general assumed a more realistic experience, since they could see different people reading the texts - this seems to be related with intelligibility. Other reasons that could have affect these two interactions regards the implementation choices. Since these are embodied interactions, it was assumed that there needed to be a degree of latency, to adjust to sudden movements when the users were trying to find the pairs in the scene. A one second delay was introduced to prevent "selection noise" when the users were trying to listen to sources that were not adjacent. Again,

the feedback proven that this implementation was in fact a constrain when the participants needed a quick adaptation and to "jump" between sources. To add up to that, the eye tracker induced quite a lot of "noise" since it capture the constant movements of the eyes, which are never in a static position.

For the second experiment, the goal was to bring the same subjects that participated in the first experiment and ask them to perform again the pairing task with the calibration they previously defined during the virtual cubes scenario. In this way, bringing a realistic environment, the goal was to understand if the virtual prototyping could provide any enhancements of the real world, compared to a natural listening condition where no interaction tools exist, and we depend on our physical and cognitive abilities to make sense of the sounds around us. Though the results shown no significant difference, possibly because the sample size was indeed quite small, there are some striking differences in the scores for the three interactions when compared with the task scores in the first experiment. As seen in the table 5.1, the participants shown an impressive improvement when performing in a realistic scenario, particularly with the head interaction, which become the most efficient one in the second experiment. Again, the reinforce of the visual field might have influenced the use of this interaction, as participants could face a certain person and attempt to lip read. In this sense, an extra tool to design the best strategy to complete the task was presented and thus the experiment also tended to have a visual-feedback condition.

Mean values of the interactions' scores		
Interaction Model	**Experiment 1**	**Experiment2**
Hand Controller	3	3
Head Rotation	1	4
Eye Tracker	1	3

Table 5.1: Average of scores for each interaction model, between scenarios

This visual-biased condition might have influence the results, however, when asked to give an opinion about their experience in comparison with the normal hearing condition (where no interaction and no calibration was working), participants acknowledge a better sense of localization of the sources, and that the task facilitate, since their focus towards a specific source was easier to achieve.

5.2 Towards a super hearing experience: part II

We can fairly assume that the initial hypothesis, more important, the part of the hypothesis related with the findings of optimal levels of listening control is far from being presented as a possible set of tools for an augmented hearing ability or an enhancement of already existing artificial hearing systems. This is a result of

the multiple conditions that were simultaneously introduced during the designed experience. As a next step, it would be necessary to assess the conditions in isolation, and evaluate if they still maintain the significance for each of the interactions. A progressive research would introduce these parameters isolated and with possible fixed interval ranges, providing that combinations of these fixed values would be tested with different interactions and with different participants. To increase the level of complexity, the environments should also simulate different positions, distances from the listener position and different audio contents.

Regarding the possible parameters of control, it still seems that calibration behavior of the participants shows a necessity to improve the signal-to-noise ratio (see plot 5.1: in dark blue is the average calibration between all participants, light blue and green are the best performances and orange and gray lines represent the worst scores) in terms of focused source and rejected signals and in this sense, according to [40] and [6], binaural separation of the sources through spatial cues and intelligibility - associated with the release from masking effect - can sill be described as products of directivity patterns and sound pressure level. It would again be required to perform the pairing task with just fixed values in gain and with no directivity influence (omnidirectional pattern) and vice-versa, testing different directivity patterns and understand with which ones, participants would feel more comfortable and perform better scores.

Looking at the usability questionnaires, it could be plausible to assume that product-wise, the directivity parameters do not present themselves as the most consumer-user friendly attributes of control over the sonic interactions. Even though its relevance in the experiment was rejected, it still important to test this same parameter without the possibility of controlling the sound pressure level. This natural behavior of increasing the gain value over directivity was a general tendency in all participants and it might be related with two factors: the first acknowledges that in the interaction implementation, gain has an immediate and direct effect over the source (it becomes LOUDER or quiet!) that the participant wants to listen to; on the contrary the directivity pattern only becomes perceptually recognized when the user moves its head away or towards the source, since the shape of the sound propagation is changed except at the relative θ angle of 0 degrees (right in front). The second factor that could have influence this poor use of the directivity values is the absence of room reverberation, which implies that the signals are to a certain extent directional by nature. In the presence of a reverberant space, directional cues would improve the localization of the source, and thus directivity would probably influence the listener's sensibility to the position of the source [40]. Hence, for a more user friendly product and intuitive design experience, it seems correct to assume that perhaps directivity could be introduced as a correlation of some sort of general aspect of listening such as a balance between background noise and meaningful signal, as a SNR parameter, and it would be up

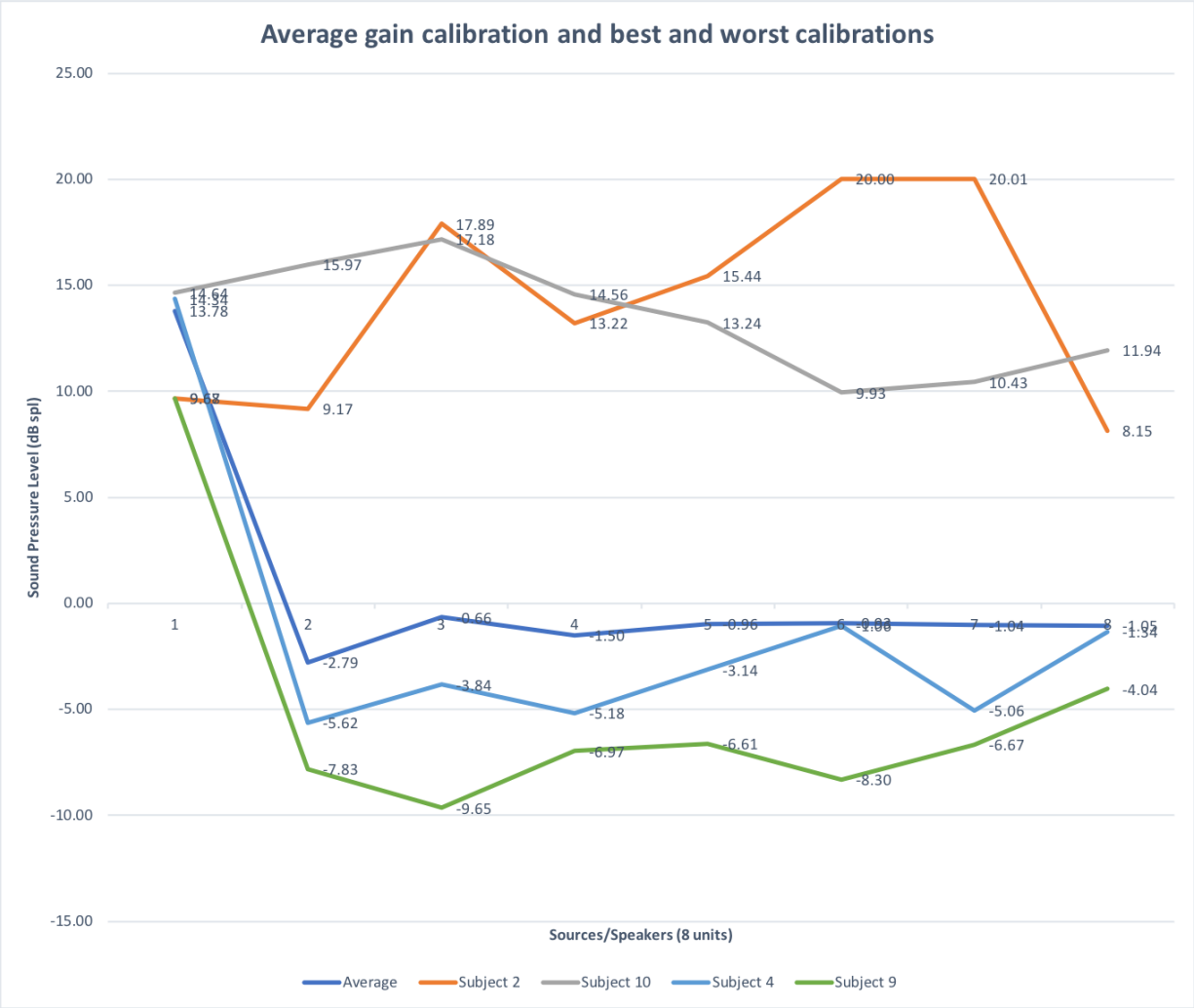


Figure 5.1: Average gain calibration (dark blue), best scores (light blue and green lines), worst scores (gray and orange lines)

to the technology to adjust the directivity beams that could achieve this optimal ratio. Again, all these assumptions would require again a larger sample group with possible new material that could involve ambisonics recordings with 360 videos and a further detailed research as described above.

The final two points of interpretation relate to the interaction that was rated as the most innovative and exiting - the eye interaction - and an interesting pattern spotted on the way calibration was performed in a general fashion.

As previously stated in Fundamentals, studies by Kimura [25] found what is commonly known as the right-ear advantage (REA) which reveals the direct

anatomic connection of the right ear to the left hemisphere, which in most people is specialized in language processing. When plotting the average calibration levels of the sources in the first experiment, it is somehow possible to see that there is an emphasis on the sources located at the right side of the fixed focused source (the one right in front at the beginning of the experiment) during the calibration, with a linear decrease of the gain value from right to left. (see figure 5.2

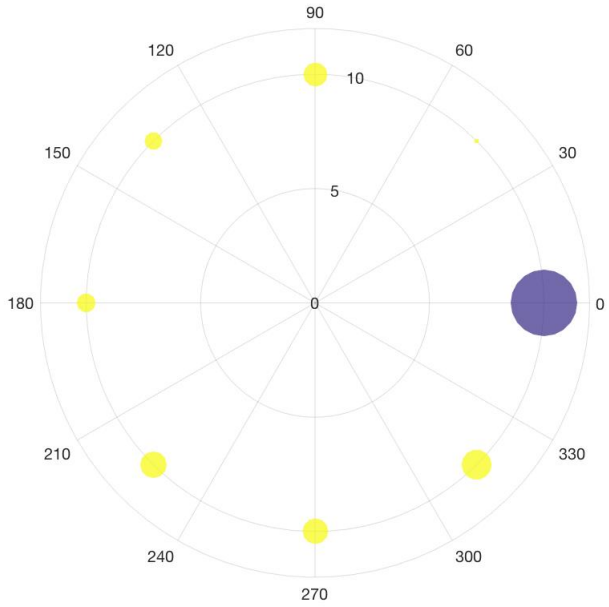


Figure 5.2: Average of sound pressure level calibrated in all interactions and represented as circles for each source in the space. The radius of the circle corresponds to the intensity of the source. The purple circle corresponds to the source in front of the listener position, which is the center of the plot. The values were normalized for a range of 0 to 1 but are the same as the ones presented in figure 5.1. The circle to the left of the purple one does not exist because it corresponds to a 0 radius circle and also to the lower value of calibration -3.9 dB FS. This behavior is verified in all interactions

Eye interaction shown to be a bittersweet metaphor, promising a very embodied and natural interaction, since the participant could easily swap between to speakers effortlessly. However, the scores for both experiments were never very high and the interaction didn't show the best performance of the three interactions. However, this could be justified by the implementation that could perhaps indulge less noise from the eye gazing (which comes already with the Pupil Labs Unity API), and on the other side, the use of only a two-dimensional coordinate system, could also reduce some of the spatial resolution. In this sense, a future implementation could also introduce a third dimension, depth and perhaps even heat mappings, to analyze counters between objects.

Chapter 6

Conclusion

This study investigates the potential use of virtual environments to develop and reproduce meaningful sonic interactions, that could help understand to what extent the listening process can be enhanced or even modulated by a set of parameters/functions. The goal of this virtual prototyping of the listening experience is to further realize if there's any significant data that would allow the creation of a super hearing system, capable of supporting the natural cognitive capacities of the auditory system.

Three models of sonic interactions were developed to facilitate the use and control of a directivity: the hand controller interaction, which works as a pointer, the head rotation interaction, that follows the frontal direction of the subject's head, and the eye tracking interaction, which tracks the eye gaze and the point in space to where the subject is looking into. These three models were tested with seventeen participants in two distinct experiments. From those seventeen, only seven took part of a second experiment, which was designed to have 360° videos. The ultimate goal of the experiment was to find the precision of the participants given these three models when asked to find which voices, in a set of eight speakers, were simultaneously reading the exact same order of sentences, between a completely virtual rendering environment where the voices were represented by cubes and the 360° video scenarios which entailed a more realistic context, with real people reading the sentences. To help the participants solve the task, two dynamic parameters of control over the directivity and sound pressure level of each source were given during the training/calibration period.

This simulation of a cocktail party scenario proven to be a stress-limit situation, where several conditions where controlled at the same time. In this sense the data results are to be taken as a first look into possible behaviors and patterns. Nevertheless, we can state that the hand controller interaction proven to be the most successful model between scenarios, being that the head rotation interaction was the one with which participants scored better results when in the presence of

360° visuals. In this sense, a more embodied interaction seems to be more efficient when the subject is surrounded by realistic stimuli and reinforced interactions, such as the hand controller, suggests a more game-like approach which brings subjects a more flexible understanding of the possible cognitive strategies and a faster feedback from their actions. Regarding the second premises of evaluation of the systems, there are significant conclusions in the correlation between the average values of calibration for each source and the best scores, but again, it is important to be able to isolate conditions to provide more solid answers.

Revisiting the research question defined in the Introduction *How the user will experience a listening situation in a realistic environment with such systems.*, it is not possible to provide a straight and single answer to this problem, but there are paths in the research of virtual prototyping - both in the sense of creating even more controlled environments as well as in more realistic situations - that could provide with more significant values: from the perspective of the user with more personalized settings but also from the perspective of context experience - each situation might required a different configuration of the parameters and the way the sonic interaction works -, improved interactions with combination of interactions, more detailed implementation - slight movements corresponds to sligh readjustments of directivity and sound pressure controls -, and even possibly a set of tools that could define a new type of listening experience, of super hearing experience.

For the next step, a paper submission is been taken place with more data collection and deeper analysis for the ACM CHI Conference on Human Factors in Computing System CHI 2019, as the topic of super hearing and virtual prototyping for sonic interactions has shown little research.

Bibliography

[1] L. L. Beranek. "Acoustical Measurements". In: *Journal Acoustical Society of America* (1949).

[2] V. Best et al. "Object continuity enhances selective auditory attention". In: *Proc. Natl. Acad. Sci. U.S.A.* (2008).

[3] J. Blauert. "Spatial hearing". In: *MIT Press* (1997).

[4] A. S. Bregman. "Auditory Scene Analysis: The Perceptual Organization of Sound". In: *The Journal of the Acoustical Society of America* (1990).

[5] D. E. Broadbent. "The role of auditory localization in attention and memory span". In: *J Exp Psycho* (1954).

[6] Adelbert W. Bronkhorst. "The Cocktail Party Phenomenon: A Review of Research on Speech Intelligibility in Multiple-Talker Conditions". In: *Acta Acustica united with Acustica* (2000).

[7] D. Byrne, H. Dillon, and K. Tran. "An international comparison of long-term average speech spectra". In: *The Journal of the Acoustical Society of America* (1994).

[8] H. Byrnes. "The Role of Listening Comprehension: A Theoretical Base". In: *Foreign Language Annals* (1984).

[9] 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* (1953).

[10] J. Christensen and J. Hald. *Beamforming*. Bruel and Kjaer Technical Review, 2004.

[11] W. T. Chu and A. C. C. Warnock. "Detailed Directivity of Sound Fields Around Human Talkers". In: *National Research Council Canada* (2002).

[12] J. F. Culling and Q. Summerfield. "Perceptual separation of concurrent speech sounds: absence of across-frequency grouping by common interaural delay". In: *The Journal of the Acoustical Society of America* (1995).

[13] J. Daniel. "Spatial Sound Encoding Including Near Field Effect: Introducing Distance Coding Filters and a Viable, New Ambisonic Format". In: *AES 23rd International Conference* (2003).

[14] J. Daniel, S. Moreau, and R. Nicol. "Further investigations of high-order Ambisonics and wavefield synthesis for holophonic sound imaging". In: *AES 114th Convention, Audio Engineering Society* (2003).

[15] H. Dillon. "Hearing Aids". In: *Boomerang Press, Sydney* (2001).

[16] Musiek. F. and M. Pinheiro. "Dichotic speech tests in the detection of central auditory dysfunction". In: *Assessment of central auditory dysfunction – Foundations and clinical correlates* (1985).

[17] Sandraand G. Hart and Lowell E. Staveland. "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research". In: (2004). DOI: 10.1016/S0166-4115(08)62386-9.

[18] V. Hamacher et al. "Signal Processing in High-End Hearing Aids: State of the Art, Challenges, and Future Trends". In: *EURASIP Journal on Applied Signal Processing* (2005).

[19] "IEEE Recommended Practice for Speech Quality Measurements". In: *IEEE Transactions on Audio and Electroacoustics* 17 (1969), pp. 225–246. DOI: 10.1109/TAU.1969.1162058.

[20] A. Ihlefeld and B. Shinn-Cunningham. "Disentangling the effects of spatial cues on selection and formation of auditory objects". In: *The Journal of the Acoustical Society of America* (2008).

[21] John C L Ingram. *Neurolinguistics: an introduction to spoken language processing and its disorders*. Cambridge University Press, 2007. ISBN: 978-0-521-79640-8.

[22] D. H. Johnson and D. E. Dudgeon. *Array Signal Processing: Concepts and Techniques*. Simon and Schuster, Inc, 1992.

[23] J. M. Kates. "Superdirective arrays for hearing aids". In: *Journal Acoustical Society of America* (1993).

[24] Gavin Kearney and Tony Doyle. "A HRTF Database for Virtual Loudspeaker Rendering". In: *The Journal of the Acoustical Society of America* (2015).

[25] D. Kimura. "Cerebral dominance and the perception of verbal stimuli". In: *Canada J. Psychol.* (1961).

[26] R. Klumpp and H. Eady. "Some measurements of interaural time difference thresholds". In: *The Journal of the Acoustical Society of America* (1956).

[27] Gunther R. Kress. *Multimodality: A Social Semiotic Approach to Contemporary Communication*. Taylor & Francis, 2010.

[28] M. Kronlachner and F. Zotter. “Spatial transformations for the enhancement of Ambisonic recordings”. In: *2nd Int. Conf. on Spatial Audio (ICSA)* (2014).

[29] H. Lane and B. Tranel. “The Lombard sign and the role of hearing in speech”. In: *Journal of Speech Hearing Research* (1971).

[30] Pierre Lecomte et al. “On the Use of a Lebedev Grid for Ambisonics”. In: *The Journal of the Acoustical Society of America* (2015).

[31] L. H. Loiselle et al. “Using ILD or ITD Cues for Sound Source Localization and Speech Understanding in a Complex Listening Environment by Listeners With Bilateral and With Hearing-Preservation Cochlear Implants”. In: *Journal of Speech, Language and Hearing Research* (2016).

[32] A. W. Mills. “On the Minimum Audible Angle”. In: *The Journal of the Acoustical Society of America* (1958).

[33] H.G. Muelle and M. Killion. “An easy method for calculating the articulation index”. In: *Hearing Journal* (1990).

[34] F. E. Musiek and G. D. Chermak. *The Human Auditory System*. Elsevier, 2015.

[35] Anja Naumann and Jörn Hurtienne. “Benchmarks for intuitive interaction with mobile devices”. In: *Proceeding in MobileHCI ’10 Proceedings of the 12th international conference on Human computer interaction with mobile devices and services*. New York: Association for Computing Machinery, 2010, pp. 401–402.

[36] M. Noisternig et al. “A 3D Ambisonic based Binaural Sound Reproduction System”. In: *AES 24th International Conference on Multichannel Audio* (2012).

[37] T. A. Powers and V. Hamacher. “Three microphone instrument is designed to extend benefits of directionality”. In: *Hearing Journal* (2002).

[38] Henning Puder. “Hearing Aids: An Overview of the State-of-the-Art, Challenges, and Future Trends of an Interesting Audio Signal Processing Application”. In: *6th International Symposium on Image and Signal Processing and Analysis* (2009).

[39] J. M. Pumford et al. “Speech recognition with in-the-ear and behind-the-ear dual-microphone hearing instruments”. In: *Journal of the American Academy of Audiology* (2000).

[40] T. A. Ricketts. *Directional Hearing Aids*. Westminster Publications, 2001.

[41] T. A. Ricketts. “The impact of head angle on monaural and binaural performance with directional and omnidirectional hearing aids”. In: *Ear Hear* (2000).

[42] T. A. Ricketts and S. Dhar. “Comparison of performance across three directional hearing aids”. In: *Journal of the American Academy of Audiology* (1999).

[43] T. A. Ricketts, G. Lindley, and P. Henry. “Impact of Compression and Hearing Aid Style on Directional Hearing Aid Benefit and Performance”. In: *Ear and Hearing* (2001).

[44] Martin Schrepp. “User Experience Questionnaire Handbook”. In: (2015). DOI: 10.13140/RG.2.1.2815.0245.

[45] B. G. Shinn-Cunningham and V. Best. “Selective attention in normal and impaired hearing”. In: *The Journal of the Acoustical Society of America* (2008).

[46] S. Stefania et al. “Sonic Interactions in Virtual Reality: State of the Art, Current Challenges, and Future Directions”. In: *IEEE Computer Graphics and Application* (2018).

[47] S. S. Stevens and Newman E. B. “The localization of actual sources of sound”. In: *American Journal of Psychology* (1936).

[48] M. Valente. *Use of Microphone Technology to Improve User Performance in Noise, Singular Publishing Group*. The textbook of hearing aid amplification, 2000.

[49] L. Vandergrift. “Listening: theory and practice in modern foreign language competence”. In: *Center For Languages Linguistics and Area Studies Conference* (2016).

[50] B. D. V. Veen and K. M. Buckley. “Beamforming: A Versatile Approach to Spatial Filtering”. In: *IEEE ASSP Magazine* (1988).

[51] Jakob Vennerød. “Binaural Reproduction of Higher Order Ambisonics A Real-Time Implementation and Perceptual Improvements”. MA thesis. Norway: NTNU - Trondheim, Norwegian University of Science and Technology, 2014.

[52] D. B. Ward, R. A Kennedy, and R. C. William. “FIR Filter Design for Frequency Invariant Beamformers”. In: *IEEE Signal Processing Letters* (1996).

[53] E. G. Williams. *Fourier acoustics: Sound radiation and nearfield acoustical holography*. Academic Press, 2005.

[54] N. Wood and N. Cowan. “The Cocktail Party Phenomenon Revisited: How Frequent Are Attention Shifts to One’s Name in an Irrelevant Auditory Channel?” In: *Journal of Experimental Psychology Learning Memory and Cognition* (1995).

[55] W. A. Yost, R. H. Dye, and S. Sheft. “A simulated “cocktail party” with up to three sound sources”. In: *Perception and Psychophysics* (1996).

Appendix A

Appendix

A.1 Experiment Material

A.1.1 Test instructions

The instructions that were read for each participant before the experiment.

First instructions

In this experiment, I want you to focus on your listening skills. By this, I mean your ability to follow or focus on a specific sound source, even in the middle of a confusing/noisy environment. There are 3 training periods and 3 tests. For each test, there will be a questionnaire about the interaction you've used. For each training and testing scene, you will be placed in the center of a circle with 8 individual cubes that are playing a set of sentences simultaneously. There are 4 cubes with male voices and another 4 with female voices.

During the experiment, you will use 3 types of interaction. And each of the 3 training sets corresponds to 1 of these 3 interaction tools, so you can first understand how to use them, individually.

There is the hand controller that works like a pointer. The headset that tracks which cube you are facing. and an eye tracker that follows which cube you are looking to.

For now, I just need to explain you:

- The controller commands
- The eye tracking calibration

Before training

There are 3 training periods and 3 tests. For each test, there will be a brief questionnaire about the interaction. For each training and testing scene, you will be placed in the center of a circle with 8 individual cubes that are playing a set of

sentences simultaneously. There are 4 cubes with male voices and another 4 with female voices.

While training, you have the chance to adjust the directivity and the sound level of each cube individually by using the hand controller, just as in the tutorial. The cubes also have a number on top of them from 1 to 8. And cube number 1 will be yellow. This is to give you a reference point.

What I want you to do now is to adjust the directivity and sound level of each cube so that you have an emphasis on cube 1, on what is it saying, without completely losing the context that there are other cubes around you - I don't want you to mute the other cubes basically. When you finish adjusting those parameters, you can move for testing. If the values prove to not be good enough for you, you can always change them until you think they are good. When testing, you will be using the values that you defined in the training for that interaction and the goal is to find which cubes are saying exactly the same sentences.

A.1.2 Sentences list and sentences' sets

Sequence 1

Set 1.1

- The birch canoe slid on the smooth planks
- Glue the sheet to the dark blue background.
- It's easy to tell the depth of a well.
- These days a chicken leg is a rare dish.
- Rice is often served in round bowls.
- The juice of lemons makes fine punch.
- The box was thrown beside the parked truck.

Set 1.2

- The hogs were fed chopped corn and garbage.
- Four hours of steady work faced us.
- Large size in stockings is hard to sell.
- A king ruled the state in the early days.
- The ship was torn apart on the sharp reef.
- Sickness kept him home the third week.
- The wide road shimmered in the hot sun.

Set 1.3

- The lazy cow lay in the cool grass.
- Lift the square stone over the fence.
- The rope will bind the seven books at once.
- Hop over the fence and plunge in.
- The friendly gang left the drug store.
- Mesh wire keeps chicks inside.
- The empty flask stood on the tin tray.

Set 1.4

- A speedy man can beat this track mark.
- He broke a new shoelace that day.
- The coffee stand is too high for the couch.
- The urge to write short stories is rare.
- The pencils have all been used.
- The pirates seized the crew of the lost ship.
- We tried to replace the coin but failed.

Sequence 2

Set 2.1

- Glue the sheet to the dark blue background.
- Rice is often served in round bowls.
- The juice of lemons makes fine punch.
- It's easy to tell the depth of a well.
- These days a chicken leg is a rare dish.
- The birch canoe slid on the smooth planks
- The box was thrown beside the parked truck.

Set 2.2

- Sickness kept him home the third week.
- A king ruled the state in the early days.
- Large size in stockings is hard to sell.
- Four hours of steady work faced us.
- The ship was torn apart on the sharp reef.
- The hogs were fed chopped corn and garbage.
- The wide road shimmered in the hot sun.

Set 2.3

- Hop over the fence and plunge in.
- Lift the square stone over the fence.
- The empty flask stood on the tin tray.
- Mesh wire keeps chicks inside.
- The rope will bind the seven books at once.
- The lazy cow lay in the cool grass.
- The friendly gang left the drug store.

Set 2.4

- The urge to write short stories is rare.
- He broke a new shoelace that day.
- A speedy man can beat this track mark.
- The pirates seized the crew of the lost ship.
- The pencils have all been used.
- We tried to replace the coin but failed.
- The coffee stand is too high for the couch.

Sequence 3

Set 3.1

- The juice of lemons makes fine punch.
- These days a chicken leg is a rare dish.
- The birch canoe slid on the smooth planks
- The box was thrown beside the parked truck.
- Glue the sheet to the dark blue background.
- Rice is often served in round bowls.
- It's easy to tell the depth of a well.

Set 3.2

- The hogs were fed chopped corn and garbage.
- Four hours of steady work faced us.
- A king ruled the state in the early days.
- The wide road shimmered in the hot sun.
- Large size in stockings is hard to sell.
- The ship was torn apart on the sharp reef.
- Sickness kept him home the third week.

Set 3.3

- The lazy cow lay in the cool grass.
- Lift the square stone over the fence.
- The friendly gang left the drug store.
- Hop over the fence and plunge in.
- The empty flask stood on the tin tray.
- The rope will bind the seven books at once.
- Mesh wire keeps chicks inside.

Set 3.4

- He broke a new shoelace that day.
- The coffee stand is too high for the couch.
- The pirates seized the crew of the lost ship.
- A speedy man can beat this track mark.
- We tried to replace the coin but failed.
- The pencils have all been used.
- The urge to write short stories is rare.

Sequence 4

Set 4.1

- The box was thrown beside the parked truck.
- The juice of lemons makes fine punch.
- The birch canoe slid on the smooth planks
- It's easy to tell the depth of a well.
- Glue the sheet to the dark blue background.
- These days a chicken leg is a rare dish.
- Rice is often served in round bowls.

Set 4.2

- Four hours of steady work faced us.
- Sickness kept him home the third week.
- The wide road shimmered in the hot sun.
- The hogs were fed chopped corn and garbage.
- Large size in stockings is hard to sell.
- The ship was torn apart on the sharp reef.
- A king ruled the state in the early days.

Set 4.3

- The lazy cow lay in the cool grass.
- Mesh wire keeps chicks inside.
- The empty flask stood on the tin tray.
- The friendly gang left the drug store.
- Lift the square stone over the fence.
- The rope will bind the seven books at once.
- Hop over the fence and plunge in.

Set 4.4

- The urge to write short stories is rare.
- We tried to replace the coin but failed.
- A speedy man can beat this track mark.
- He broke a new shoelace that day.
- The coffee stand is too high for the couch.
- The pirates seized the crew of the lost ship.
- The pencils have all been used.

Sequence 5

Set 5.1

- The birch canoe slid on the smooth planks
- The box was thrown beside the parked truck.
- It's easy to tell the depth of a well.
- Rice is often served in round bowls.
- Glue the sheet to the dark blue background.
- The juice of lemons makes fine punch.
- These days a chicken leg is a rare dish.

Set 5.2

- A king ruled the state in the early days.
- Large size in stockings is hard to sell.
- The wide road shimmered in the hot sun.
- Sickness kept him home the third week.
- The hogs were fed chopped corn and garbage.
- Four hours of steady work faced us.
- The ship was torn apart on the sharp reef.

Set 5.3

- Lift the square stone over the fence.
- The empty flask stood on the tin tray.
- The lazy cow lay in the cool grass.
- Hop over the fence and plunge in.
- The rope will bind the seven books at once.
- The friendly gang left the drug store.
- Mesh wire keeps chicks inside.

Set 5.4

- He broke a new shoelace that day.
- The coffee stand is too high for the couch.
- The pencils have all been used.
- We tried to replace the coin but failed.
- The urge to write short stories is rare.
- A speedy man can beat this track mark.
- The pirates seized the crew of the lost ship.

Sequence 6

Set 6.1

- The box was thrown beside the parked truck.
- Glue the sheet to the dark blue background.
- Rice is often served in round bowls.
- It's easy to tell the depth of a well.
- These days a chicken leg is a rare dish.
- The birch canoe slid on the smooth planks
- The juice of lemons makes fine punch.

Set 6.2

- Large size in stockings is hard to sell.
- A king ruled the state in the early days.
- The hogs were fed chopped corn and garbage.
- Four hours of steady work faced us.
- The wide road shimmered in the hot sun.
- The ship was torn apart on the sharp reef.
- Sickness kept him home the third week.

Set 6.3

- Hop over the fence and plunge in.
- Mesh wire keeps chicks inside.
- The lazy cow lay in the cool grass.
- The empty flask stood on the tin tray.
- The rope will bind the seven books at once.
- The friendly gang left the drug store.
- Lift the square stone over the fence.

Set 6.4

- The urge to write short stories is rare.
- The pirates seized the crew of the lost ship.
- We tried to replace the coin but failed.
- The coffee stand is too high for the couch.
- A speedy man can beat this track mark.
- He broke a new shoelace that day.
- The pencils have all been used.

Sequence 7

Set 7.1

- The birch canoe slid on the smooth planks
- Glue the sheet to the dark blue background.
- The box was thrown beside the parked truck.
- It's easy to tell the depth of a well.
- The juice of lemons makes fine punch.
- These days a chicken leg is a rare dish.
- Rice is often served in round bowls.

Set 7.2

- The ship was torn apart on the sharp reef.
- Sickness kept him home the third week.
- The hogs were fed chopped corn and garbage.
- A king ruled the state in the early days.
- The wide road shimmered in the hot sun.
- Large size in stockings is hard to sell.
- Four hours of steady work faced us.

Set 7.3

- The empty flask stood on the tin tray.
- Lift the square stone over the fence.
- The rope will bind the seven books at once.
- The lazy cow lay in the cool grass.
- Hop over the fence and plunge in.
- The friendly gang left the drug store.
- Mesh wire keeps chicks inside.

Set 7.4

- A speedy man can beat this track mark.
- The urge to write short stories is rare.
- He broke a new shoelace that day.
- The coffee stand is too high for the couch.
- We tried to replace the coin but failed.
- The pencils have all been used.
- The pirates seized the crew of the lost ship.

TABLE 1 : randomized order of speakers for each set

Sequence	Set	Speakers
1	1.1,1.2,1.3,1.4	1 2 8 3 7 4 6 5
2	2.1,2.2,2.3,2.4	2 3 1 4 8 5 7 6
3	3.1,3.2,3.3,3.4	3 4 2 5 1 6 8 7
4	4.1,4.2,4.3,4.4	4 5 3 6 2 7 1 8
5	5.1,5.2,5.3,5.4	5 6 4 7 3 8 2 1
6	6.1,6.2,6.3,6.4	6 7 5 8 4 1 3 2
7	7.1,7.2,7.3,7.4	7 8 6 1 5 2 4 3

TABLE 2 : Randomize sequence order:

PILOT :

Subject	Sequence	Training	Single interaction
1	1,7,3,2,6,5,4,7,5	3,2,1	3,2,1

TEST 1:

Subject	Sequence	Training	8 speakers test
1	7,5,2,4,1,6,3,7,5	2,1,3	2,1,3
2	4,1,7,5,6,3,2,6,1	1,2,3	1,2,3
3	4,7,3,1,5,6,2,4,5	3,1,2	3,1,2
4	2,7,4,3,1,5,6,3,2	1,3,2	1,3,2
5	5,4,3,7,2,6,1,4,7	1,3,2	1,3,2
6	5,7,6,4,3,2,1,6,2	1,2,3	1,2,3
7	4,3,5,6,1,2,7,6,5	3,1,2	3,1,2
8	7,3,2,5,6,4,1,4,6	3,1,2	3,1,2
9	4,2,7,6,5,3,1,4,3	3,2,1	3,2,1
10	7,2,3,1,5,6,4,1,5	3,1,2	3,1,2
11	6,2,5,1,7,4,3,1,3	1,2,3	1,2,3
12	1,2,6,3,4,7,5,4,2	3,1,2	3,1,2
13	2,3,6,5,4,1,7,6,4	1,3,2	1,3,2
14	3,6,7,5,4,1,2,1,3	3,1,2	3,1,2
15	6,1,5,3,2,4,7,1,4	2,1,3	2,1,3
16	4,2,5,6,3,1,7,3,6	3,1,2	3,1,2
17	5,1,4,6,2,7,3,4,6	1,2,3	1,2,3

TEST 2:

Subject	Sequence	Training	Single interaction
1	3,1,2,4,7,5,6,4,2	2,3,1	2,3,1
2	5,2,1,3,7,4,6,2,4	3,1,2	3,1,2
3	1,4,3,2,6,5,7,6,1	2,1,3	2,1,3
4	4,6,3,7,2,1,5,3,6	2,3,1	2,3,1
5	7,1,5,2,4,3,6,7,6	3,1,2	3,1,2
6	6,1,3,2,4,5,7,1,5	2,1,3	2,1,3
7	1,4,2,6,5,7,3,7,6	2,1,3	2,1,3