Positional Infrared Tracking Based System Using Non-individualised HRTFs to Simulate a Loudspeaker Setup and its Influence on Externalisation of Music

Master Thesis

Rasmus Eklund

Aalborg University
Sound and Music Computing

AALBORG UNIVERSITY
STUDENT REPORT

Sound and Music Computing
Aalborg University
http://www.aau.dk

Title:
Positional Infrared Tracking Based System Using Non-individualised HRTFs to Simulate a Loudspeaker Setup and its Influence on Externalisation of Music

Theme:
Sonic Interaction Research

Project Period:
Spring Semester 2019

Project Group:
Rasmus Eklund

Participant(s):
Rasmus Eklund

Supervisor(s):
Cumhur Erkut

Copies: 1

Page Numbers: 41

Date of Completion:
May 28, 2019

Abstract:

A large number of people produces and mixes their audio productions only with headphones. Consequently, because of acoustical differences including the inter-aural level and time differences, listening on loudspeakers might sound dissimilar compared to headphones. However, using Head Related Transfer Functions HRTFs, it is possible to simulate a loudspeaker setup using the binaural panning techniques. In this report an infrared based positional tracking system using non-individualised HRTFs to simulate a loudspeaker setup is conceptualised, designed and implemented. The system was evaluated on 20 participants to see if the addition of positional tracking increased the degree of externalisation. There was not found a significant difference between the only head movement and additional position conditions. Comparisons to previous studies were discussed and improvements for future experiments were proposed.

Sound and Music Computing
Aalborg Universitet
http://www.aau.dk

AALBORG UNIVERSITET
STUDENTERRAPPORT

Titel:
Positional Infrared Tracking Based System Using Non-individualised HRTFs to Simulate a Loudspeaker Setup and its Influence on Externalisation of Music

Tema:
Sonisk interaktion

Projektperiode:
Forårssemestret 2010

Projektgruppe:
Rasmus Eklund

Deltager(e):
Rasmus Eklund

Vejleder(e):
Cumhur Erkut

Oplagstal: 1

Sidetal: 41

Afleveringsdato:
28. maj 2019

Abstract:

Et stort antal mennesker producerer og mixer deres lydproduktioner kun med hovedtelefoner. På grund af akustiske forskelle, herunder de interaurale niveau og tidsforskelle, det at lytte via højttalere kan lyde ulig i forhold til hovedtelefoner. Men ved hjælp af HRTF'er er det muligt at simulere en højttaleropsætning ved hjælp af binaural panoreringsteknikker. I denne rapport er et infrarødt baseret positionssporingssystem, der bruger HRTF'er der simulerer en højttaleropsætning, konceptualiseret, designet og implementeret. Systemet blev evalueret på 20 deltagere for at se, om tilsætning af positionelle sporing øgede graden af eksternalisering. Der blev ikke fundet en signifikant forskel mellem disse to grupper. Sammenligninger med tidligere eksperimenter blev drøftet og blev foreslået forbedringer for fremtidige eksperimenter.

Contents

Preface

This is my thesis work for my Master education in Sound and Music Computing at Aalborg University in Copenhagen. It combines the skills I obtained through my Bachelor's degree in Medialogy and the Sound and Music Computing Master. The theme of this paper reflects and combines some of my great interests that I have learned throughout the education, but also from my interests in my free time. The project includes the implementation of a positional tracked system that uses non-individualised HRTF for a spatialised experience for the listener. It is implemented with the intention to simulate a loudspeaker setup through headphones and is a tool for music producers, who mainly mix their songs on headphones, get a chance to listen to how it would sound like on loudspeakers, but through headphones. The project was made possible with the supervision by Cumhur Erkurt.

Aalborg University, May 28, 2019

Rasmus Eklund

Rasmus Eklund
<reklun13@student.aau.dk>

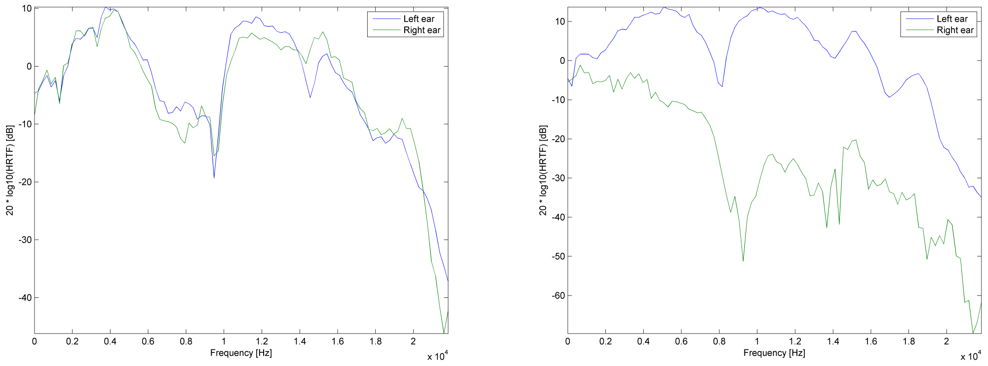
Chapter 1

Introduction

Listening to music on headphones and loudspeakers are two very different experiences. Loudspeakers propagate sound waves that reaches the listeners ears at different sound pressure levels and times from the sound source. Contrary to headphones, the left and right channel is completely separated from each ear with the two driver units located close to each ear. Also, the driver units are also always the same distance to the ear, so the level does not change if the listener changes position. Hereby, there are no acoustical effects and spectral cues affected to the listeners as compared to listening on loudspeakers [6]. A common problem that many music producers face is that a produced song sounds very different from one sound system to the other, because of these acoustical differences. The challenge can occur if the producer does not have access to a set of loudspeakers or does not have a acoustically treated room to listen to the music. It becomes more convenient to produce music solely on headphones instead, but this leaves out the possibility to experience the produced music with other sound system and/or in other environments.

Not only does the interaural time and level differences (ITD and ILD) have an influence for humans when listening to sounds, but the the shape of the outer ear (the pinna) works as a acoustic filter for humans to better localise sounds in 3D space. However, the anatomical features of humans are idiosyncratic meaning that each person has their own set of personalised filters [9].

Throughout the years, the field of binaural hearing has been investigated. The acoustic paths from the sound source to the two ears can be represented as filters and simulated using DSP algorithms in terms of the Head Related Transfer Functions (HRTFs). For each position of the sound source there is a corresponding filter for each ear. This is usually acquired with specialised equipment where microphones are placed in the persons ear and impulses are recorded at various positions around the head in an anechoic chamber as seen on figure 1.4. Therefore, to record and measure these individualised HRTFs it requires a lot of equip-



(a) **(b)**

Figure 1.1: The HRTF frequency response of a subject for a sound source placed directly ahead of the listener (a) and the sound source directly to the left (b) [27].

ment and is also a cumbersome process, but to reach maximum realism, individual HRTFs are required. Non-individual HRTFs can cause confusion and the listener can localise the position of a sound source incorrectly. This is due to the HRTFs are strongly determined by the filtering properties of the anatomy of the outer ear, head, shoulders and torso, which are idiosyncratic [16]. Therefore, by listening to non-individualised binaural recording, one may perceive the audio scene inadequately and might not perceive it as externalised. Externalisation is the distance perception that is related to binaural listening and can also be called *out-of-the-head-localisation* [6]. Normally when listening on headphones it sounds like the stimuli is coming from inside the head (internalisation), but with the use of HRTF for binaural listening, the sounds can be perceived as being outside of the head and/or in close reach. Moreover, sufficient externalisation might be achievable even though the HRTF is not personalised. Research have shown that even though using non-invidualised HRTFs, subjects are still able to accurately localise virtual sources compared to free-field sources [26].

There has been done several HRTF measurements that are used as databases for use in various scenarios. One prominent is the CIPIC database which was measured at the U. C. Davis CIPIC Interface Laboratory [2]. It includes head-related impulse responses for 45 subjects at 25 different azimuths and 50 different elevations which gives 1250 directions at approximately 5° angular increments (as seen on figure 1.3). Furthermore, the anthropocentric measurements for each subject are also included. They did also make measurements with the Knowles Electronics Manikin for Acoustic Research (KEMAR) with a large and small pinnae. This is usually used as a standard as the torso, head, and the shape and size of the ear is based of an average of about 5000 males and females [14]. The measurements

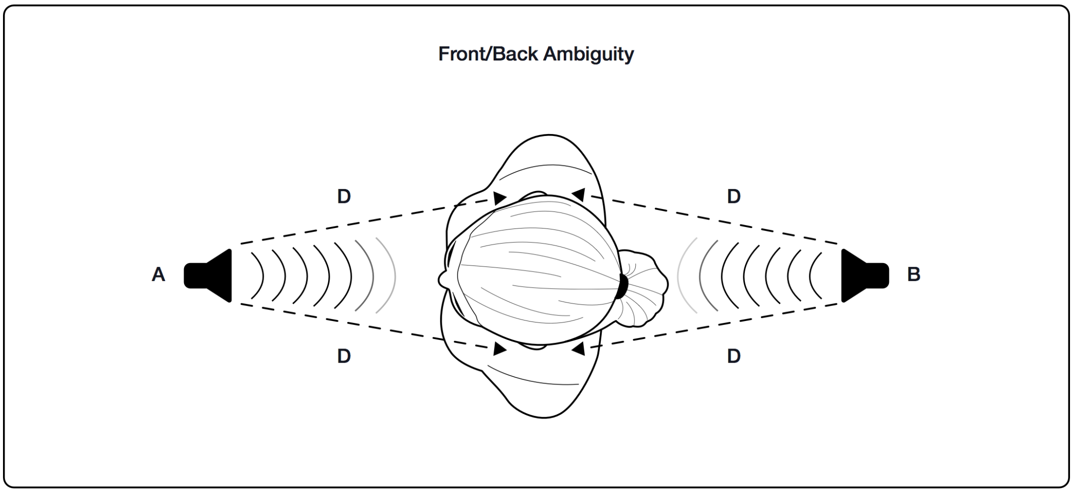


Figure 1.2: Two sound sources (A & B) with equal length (D) located at the front an rear to the listener [21].

were done by having subject seated at the center of a 1 meter radius hoop whoose axis was aligned with the subject's interaural axis. They produced a head-related impulse response (HRIR). And was recorded by probe microphones placed in close to the subject's ear canal. Each HRIR were 200 samples corresponding to 4.5 ms at a 44.1 kHz sampling rate. Given the measurements they found HRTF variation for the ITD with a corresponding ±10.3% which is also strongly correlated with the head size. On further inspection in the frequency domain most HRTFs have a prominent resonance around 3-4 kHz caused by the pinna - hence it is called the "pinna-notch frequency". This difference for the frequency response for the different HRTFs also correlates with the anatomy of the individuals ear.

Another great challenge in regards to externalisation is the front-rear confusion. As lateral sources are almost always judged to be external, frontal and rear are most likely to be perceived inside the head, or misjudged as to be frontal or rear [16]. When two sound sources are the same distance front and rear relative to the listener, one cannot rely on time or level differences as they are identical. Humans instead rely on spectral modifications caused by the head and body that create these natural filter. For example, on figure 1.2 we can see two sound sources that have equal distances to the listener's ears, but frontal sounds produce resonances created of the pinna, while rear sounds are shadowed by the pinna.

The spectral modifications by themselves may not be enough for listeners to localise a sound precisely, so we rely on head movements to assist with localisation. By simply turning the head, one can more easily distinguish between front and back sound sources [21].

Hendrickx *et al.* [16] showed that sufficiently large head movements that are coupled with head tracking can enhance externalisation for frontal and rear sound

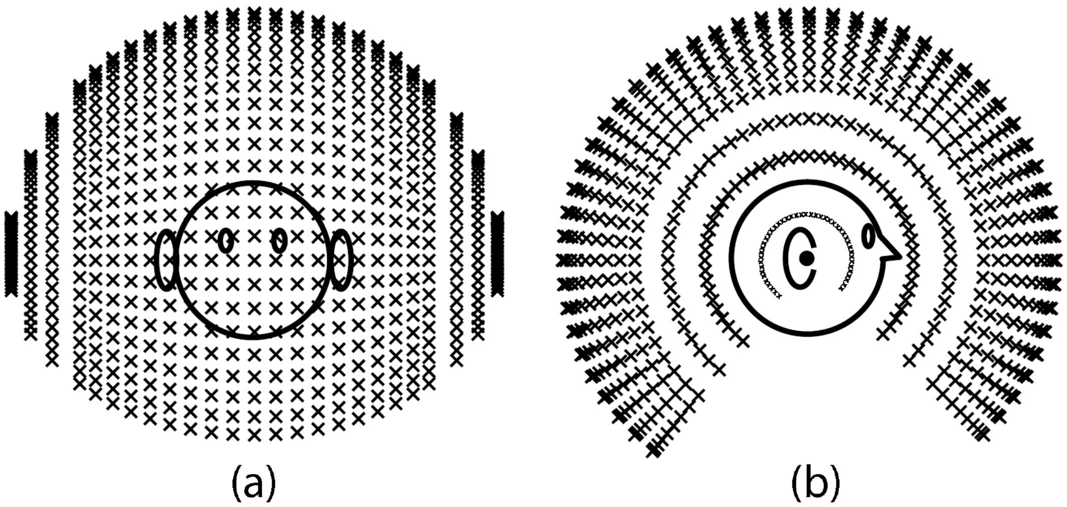


Figure 1.3: Locations of data points (a) front (b) side, used for the CIPIC HRTF database [14].

sources compared to when subjects do not move their head. Studies have shown that head movements enables subjects to localise sources more accurately also because sound sources are in constant motion in respect to the listener as the head is never perfectly still. However, this is not the case if the subject listens to binaural content through headphones without any head tracking - the location of the sound sounds moves with the listener. Their research showed that the externalisation persisted even though the subject stopped moving the head. This also confirms the study made by Brimijoin *et al.* [7] that found that when subjects slightly moved their head back and fourth between 15° there was a difference in the degree of externalisation compared to no head movement. Head-moving trials for signals that remained fixed to the world, but not to the players head movements were externalised 65% of the trials compared only 20% of the trials where the signals were fixed relative to the head.

However, the previous mentioned research only rely on head rotation with 3 degrees of freedom (DOF) and does not take the listeners position into consideration to acquire 6 DOF. Brimijoin *et al.* [7] also mentions that there is a reasonable claim that one cannot externalise sound if it's to have zero distance from the head and that there has been done very little work using motion tracking to examine distance perception. It has been shown that head movements are useful in distance perception that the sound intensity and the familiarity of the sound, the listener can quite accurately judge the distance of it [10]. If the listener also can move in 3D space and come closer or further away from the sound source, the listener would more easily judge the distance of the source and potentially increase the externalisation and realism in general. The level of a sound is the most simple way for humans to determine how far or close a sound is to the human ear - the closer the

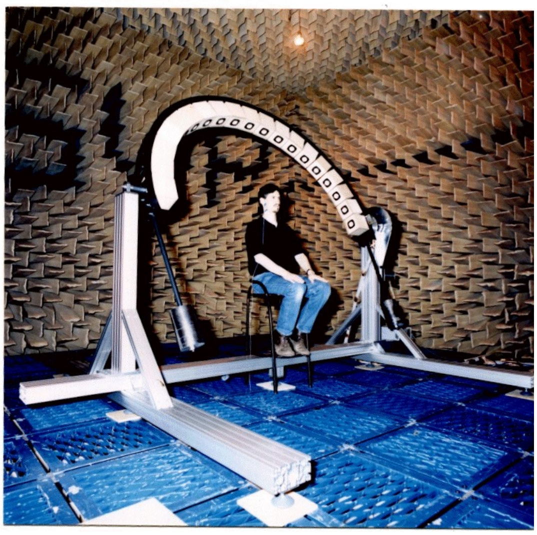


Figure 1.4: A setup of measuring HRTFs in an anechoic chamber. The arc surrounding the person is speakers at different angles and is recorded by microphones placed inside the ears of the subject. [27].

sound is the louder it is [9]. If the listener only can determine the location of the sound in terms of how far away it is by only rotating the head, it might be difficult as the extra dimension of moving to the sides and back and forth can be a crucial factor for determine the location of the sound.

Additionally, it has been shown that to achieve even further realism reverberation that matches the spectrum that of free-field signals tend to be perceived as externalised (sounds that appear to be ’out of the head’, contrary to internalised that is the case when listening on headphones) [7]. In the study of Hendrickx *et al.* the speech stimuli they used had small amount of reverberation as they mention that the externalisation rates might have been higher, whether or not the head tracking is active. Ideally, the reverberation should correspond to the room the subject is located in to exact match the realism. Furthermore, reverberation is also an essential cue for distance perception if it matches the environmental context. However, this is difficult since the correct amount of reverberation can only be accurate if it is obtained by carrying out acoustical measurements of the particular room or environment [17]. Additionally, the reverberation also changes depending if it is near-field (withon 1 meter of the listener) and the signal becomes more dry in terms of the early reflections and diffuse reverberation [5].

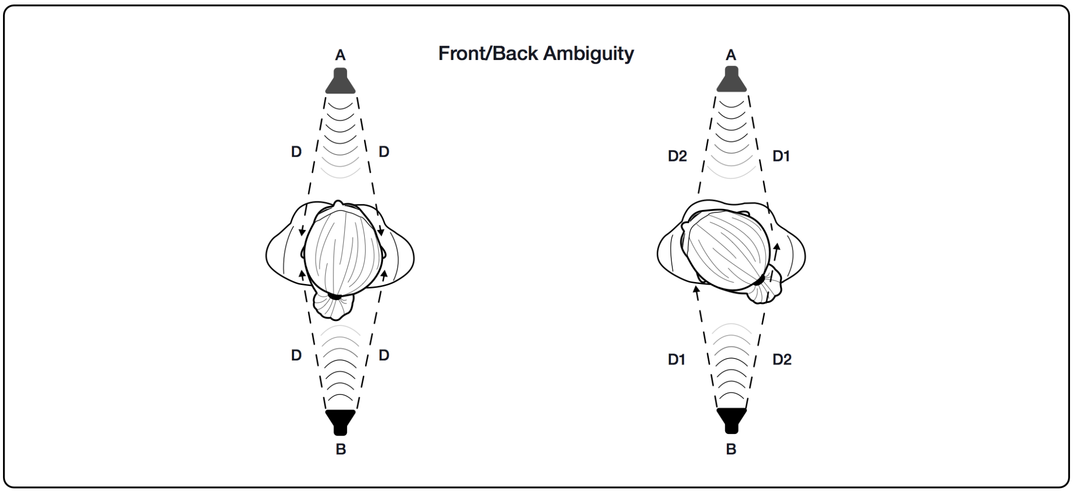


Figure 1.5: By rotating the head, the listener can shorten (D1) the distance from the sound source to the ear and at the same time lengthen it (D2). By doing so the spectral modification changes and so does the time and level differences [21].

1.1 Paper structure

The paper is structured as follows. The next chapter presents and explained the related work within Binaural hearing and the use of HRTF both for software (plug-ins) and hardware (headphones). This leads to the design requirements in chapter 3. Here a problem statement is formulated based on the two previous chapters and it leads to the design requirements. In chapter 4 a review of the design and implementation of the hardware and software of the system will be explained. In the end of the chapter a hypothesis that will be the basis of the experiment is presented. Chapter 5 presents how the experiment was set up, including choice of stimuli, location and experimental protocol. Chapter 6 shows the results of the experiment and statistical tests are made to either confirm or reject the hypothesis. Chapter 7 discusses the results of the experiment and compares it with other relevant studies and discusses potential improvements. Chapter 8 concludes the paper.

Chapter 2

Related Work

In this chapter the current technologies within HRTF and its application for simulating externalised sounds through headphones will be outlined. First, the state of the art software within spatial audio for music production will be investigated. The focus will be on plugins for digital audio workstations (DAWs) that lets the user interact and change parameters including azimuth and elevation, but some of them also includes room emulation with room reflections and reverb for a more realistic simulated experience.

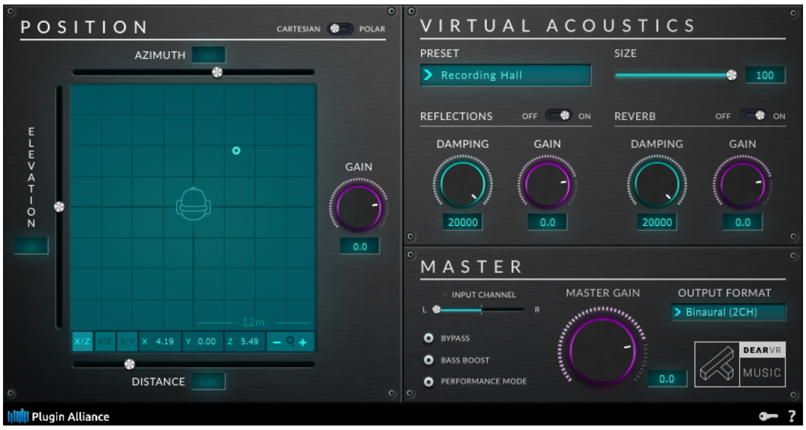
Finally, the recent field of "3D headphones" will be introduced both commercial examples and conceptualised concepts. This includes integrated head tracking and anthropometric customisation used for sound localisation and room emulation to give the user a more immersed and cinematic experience compared to what conventional headphones can offer.

2.1 Software (Plugins)

For many artists and producers, the use of binaural panning can create a more spatial audio experience for stereo projects. There are a lot of useful software solutions that makes this possible. Some focus solely on the azimuth and elevation parameters, to let the user locate a sound in 3D space. Some also includes reflection and room ambience for more precise realism. The following examples have basic implementation and others are more advanced to let the user have control of the binaural listening experience.

2.1.1 Sennheiser Ambeo orbit

The Sennheiser Ambeo Oribit lets the user place a sound source in 3D space based on its azimuth and elevation. It is based on binaural recording using the Neumann KU100, a dummy head used for binaural stereo recordings [22]. In addition to



X 4.19 Y 0.00 Z 5.49

GAIN 0.0

PRESET Recording Hall

SIZE 100

REFLECTIONS DAMPING 20000 GAIN 0.0

REVERB DAMPING 20000 GAIN 0.0

MASTER GAIN 0.0

OUTPUT FORMAT Binaural (2CH)

Figure 2.1: The DearVR that lets the user alter the position of the sound source and also the virtual acoustics including the type of room, reflections and amount of reverb.

locate the sound, it is also possible to add reflection to simulate the sound in a room with a specific size. The user can also change the material the room is made of to alter the early and late reflections of the room and to improve the spatial accuracy compared to a reverb plugin. The clarity option alters the timbre and 3D externalisation of the incoming audio.

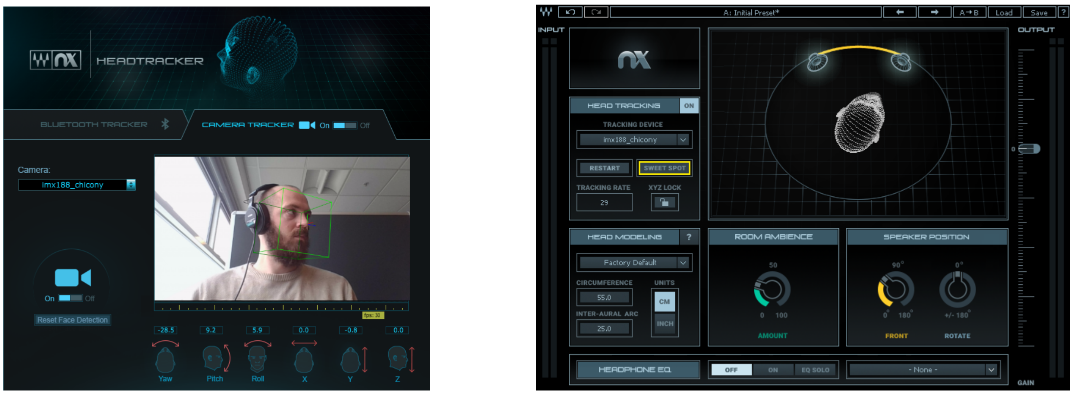
2.1.2 DearVR music

DearVR specialises in tools to create more immersive 3D audio [8]. With the DearVR Music plugin, the aim is to make a more immersive audio production while using headphones. Based on their audio reality engine, they aim to imitate the acoustic modelling of an environment that does not only focus on the spatial location, but also combines, distance, motions, reflections and reverb. In the graphical user interface, the elevation, azimuth and distance can be altered which gives additional sense of depth as the distance controls the gain of the sound source.

2.1.3 FFT-based binaral panner

An open source project that tries to create realistic 3D-audio through headphones is the "FFT-based Binaural Panner" by Jakob H. Andersen [3]. The patch is made is Cycling '74 Max. The project uses recordings from the CIPIC HRTF database to make the binaural panning. It was done to reduce the load on CPU when making convolution in the time domain, enhance the process of FFT is used to do the process in the frequency domain instead.

Since the patch is based on measurements from the CIPIC database, a "HRTF-SubjectMatcher" class is made for user to insert their own anthropometric measurements used for HRTF measurements to find the subject that matches closest to



(a) (b)

Figure 2.2: The Waves Nx plugin with the headtracker software (a) measuring pitch, yaw, roll and z, y, z based on the facial recognition software. The main software (b) simulates the head movement based on the position of some virtual loudspeakers placed in front of the listener.

a subject from the actual database. This is to more accurately match measurements and give the best matched filters to the subject. Otherwise a non-individualised HRTF set can be initialised and used as well.

One can alter the azimuth and elevation and distance to place the sound object in 3D space. It does however not include room emulation, but it emulates the distance from sound object to the listener by decreasing the gain of the sound the farther away the listener gets.

An external java object handles the direction and distance calculations based on the listener position in the XYZ plane. Additionally the rotation of the listener (unit quaternion) is also used to calculate the correct direction and distance from sound source to listener. Hereby, the sound volume and the delay for both ears (left and right channel) are calculated in real-time.

Kasper Skov has taking this further and made a Max for Live plugin based on the FFT-based binaural panner [24]. It contains the same features as the original patch made in Max, but now it has a graphical user interface to easily place sounds in 3D space and get it visualised in a virtual 3D room made with jitter.

2.1.4 Waves Nx

The Waves Nx works as a virtual room emulator over headphones. Hereby the user can monitor 7.1, 5.1 and 5.0 surround on stereo headphones [25]. The use case of the software is for producers who want to monitor mixes over headphones in case you do not have a acoustically good room or primarily mix on headphones and do not have loudspeakers available.

Contrary to the other mentioned software solutions the Waves Nx has a "head modelling" feature that let you measure your circumference and inter-aural arc to calculate the inter-aural delays, filters and gains for each ear and hence used to approximate an individual HRTF. By default average data is set for the adult human population.

Another interesting feature is the head tracking via camera. This feature tracks the orientation of your head and makes the sound stay in the same position to match a real life scenario. The camera based tracking works by a facial recognition algorithm the track the position and rotation of the face as seen on figure 2.2. The limitation of this solution is that the camera requires enough visible light to recognise the face and most cameras integrated in laptops have low frames per second, especially used in dark environments. Also, the camera based solution does not allow for a full 360 deg rotation as the facial recognition only works when the face is detected.

2.2 Hardware (Headphones)

Conventional headphones work great for binaural hearing as DSP is applied and sent to the two channels for the left and right ear. However, the term "3D headphones" has been introduced that tries to immerse and give a more cinematic experience compared to conventional headphones. In the following section some concepts and commercial 3D headphones will be presented.

2.2.1 OSSIC X

The kickstarter project "OSSIC X" is a calibrated 3D headphone for a personalized HRTF experience [18]. The idea behind the project is to make a headphone that takes the anatomy of the listener into account to make an individual HRTF. This include the size of the head and shape of the ears. By using this data, they claim to make the listening experience ten times more immersive than current technologies according to OSSIC.

They made some acoustical, physics-based testing with a microphone placed in the ear to test if the headphone reproduce the same frequency response as a real sound in space. They compared it to a generic gaming headphone and the results look promising as seen on figure 2.3. They also tested sound localisation estimation and found that people had better accuracy in regards to angle and depth estimation of locating a sound listening with the OSSIC X compared to a generic gaming headphone.

It also has integrated head tracking and anatomy calibration. However, here you do not need to specify the measurements yourself, but as the headphone is put on the head, the inter-aural arc is measured to determine a more personalised HRTF. To further enhance the realism, each earcup has 4 drivers. According to

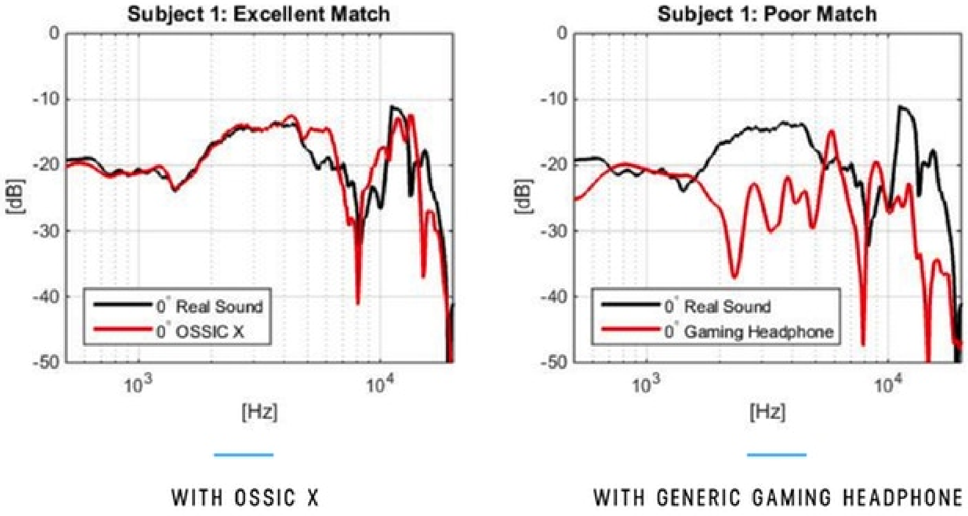


Figure 2.3: Comparison of frequency response from the OSSIC X headphone to a generic gaming headphone.

OSSIC this ensures accurate sound location playback, such as to more easily distinguish sounds at different elevations.

Unfortunately, the project got discontinued in 2018 and the headphones are no longer in production.

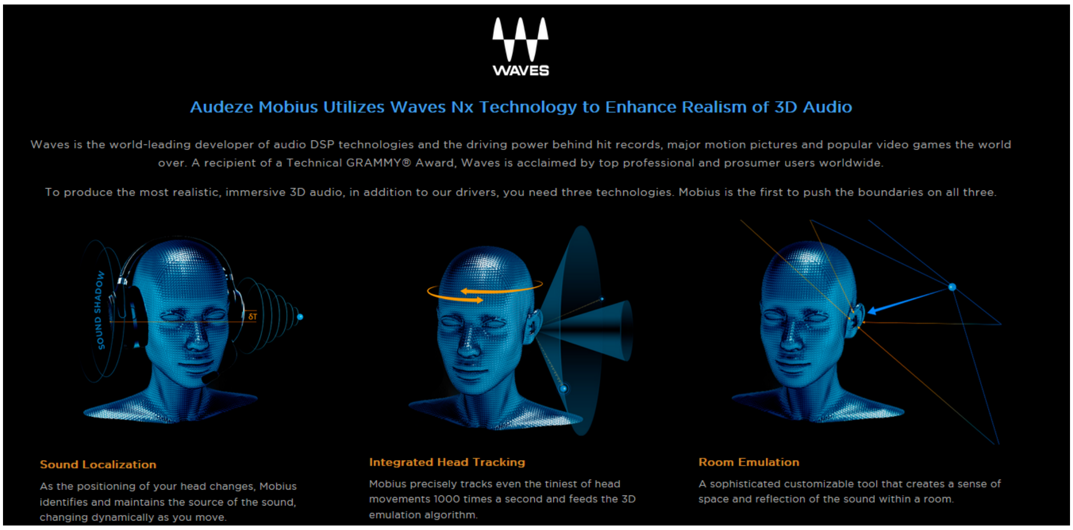
2.2.2 Audeze Mobius 3D headphones

In collaboration with Waves Nx (mentioned in 2.1.4) Audeze has produced the Mobius what claims to be the world's first premium 3D cinematic headphone to deliver realistic and immersive 3D audio. It employes the technology of the Waves Nx head tracker and is integrated in the headphone. It also has anatomy calibration for a estimated individualized HRTF. This includes a lot of the features that the Ossic X also offered and the Audeze Mobius is currently commercially available.

Just like with Waves Nx, the Audeze Mobius is included with software that let you insert your head circumference and inter-aural arc for HRTF personalisation.

2.2.3 Sennheiser Ambeo smart headset

Contrary from the two previous examples, the Sennheiser Ambeo smart headset is an in-ear headset that can do binaural recording [23]. Two miniature microphones are placed inside each earpiece. These microphones record the 3D soundscape by utilizing the physical structure of the outer ear as one world hear naturally. In a sense the headset can capture an individualized HRTF. The use of microphones



Audeze Mobius Utilizes Waves Nx Technology to Enhance Realism of 3D Audio

Waves is the world-leading developer of audio DSP technologies and the driving power behind hit records, major motion pictures and popular video games the world over. A recipient of a Technical GRAMMY® Award, Waves is acclaimed by top professional and prosumer users worldwide.

To produce the most realistic, immersive 3D audio, in addition to our drivers, you need three technologies. Mobius is the first to push the boundaries on all three.

Sound Localization

As the positioning of your head changes, Mobius identifies and maintains the source of the sound, changing dynamically as you move.

Integrated Head Tracking

Mobius precisely tracks even the tiniest of head movements 1000 times a second and feeds the 3D emulation algorithm.

Room Emulation

A sophisticated customizable tool that creates a sense of space and reflection of the sound within a room.

Figure 2.4: The Audeze Mobius utilizes the Waves Nx Technology in three different aspect of sound localisation, integrated head tracking and room emulation.

also offers active noise cancelling and 'Transparent Hearing', whereby the signal from the microphones are passed to the earpieces.

2.3 Summary

In this chapter various state of the art technology within binaural hearing, HRTF and spatialised audio for both software and hardware purposes have been discovered. There are currently a lot of software solutions that uses HRTF to do spatialised sound where the user can change parameters such as the elevation and azimuth and room emulation and/or reverb. Waves Nx are one of the few examples that includes head tracking to automatically determine the position and rotation of the listeners head and thereby in real-time update the parameters for an immersive 3D audio experience.

Audeze has utilised this head tracking technology and integrated that in their new Mobius headphones. But this still only uses the head rotation as part of the tracking technology and the position is therefore not tracking in the case. The distance from the sound source to the listener is not part of the integrated technology, but the room emulation feature is what is an alternative to have a perceived distance from the listener to the sound source. This could therefore be an point of interest to further investigate the implementation of position tracking that can be used to simulate a loudspeaker setup. Furthermore, it would be intriguing to examine how listeners perceive this simulated setup and if the sound sources can be externalised to some extent.

Chapter 3

Design Requirements

It is fair to argue that there is a missing piece in the research on the externalisation of sound when both head movement and position is tracked. Since the position of the listener relative to the sound source is a great factor for the degree of externalisation, it seems natural to have this implemented in the proposed system. In the current research the focus has merely been on the head movement in terms of externalisation and the actual position has not been a focus point, even though this is a fundamental way for humans to localise sounds [9]. Head movement is still a very crucial way for humans to localise sounds, but the addition of the position where the loudness of the sound source alters depending on the distance relative to the listener, might increase the rate of externalisation.

There has been an outline of the state of the art technology within spatialised audio both for software that uses HRTFs to simulate binaural sound and the new phenomena of 3D headphones with integrated head tracking. It has also introduced how HRTFs (both individualised and non-individualised) can be used to simulate a sound in 3D space while listening on headphones and how head movements coupled with head tracking can improve the degree of externalisation. Therefore, a problem statement is created to support the decisions made in the design requirements and the experiment that follows:

"How can a headphone based system simulate a loudspeaker setup by using a positional tracking system with non-individualised HRTFs? And can this system further improve the degree of externalisation compared to only using head movement tracking?"

Given this recent technology and the research within academia for HRTF and the degree of externalisation with headphones, design requirements have been made for the purposed system. It is divided into two sections (software and hardware part) for a more detailed explanation and the reasoning for these requirements. At the end of the chapter, the design requirements are listed in short.

3.1 Software: FFT-based binaural panner with non-invidualised HRTFs

The sound processing of the system will be based on the FFT-based binaural panner by Jakob H. Andersen [3]. As a starting point this implementation works great for azimuth and elevation based on the CIPIC HRTF database. In order for the system to work for the problem statement, two sound sources (left and right loudspeaker) should be present and stationary and then the sound receiver (listener) that can move and rotate and the appropriate angle and distance to the sound sources should be calculated.

It is also decided to use the CIPIC HRTF dataset and hence the HRTFS will not be individualised to the listener. The reason for doing this is because the system is meant to be an easy and accessible tool without too much configuration and calibration. With the current solutions for making individualised HRTFS, it will be a long and cumbersome process to gather the information and to measure. Furthermore, several studies has been using non-individualised HRTFs for their experiments and had similar results compared to the ones that used individualised HRTFs. The only noticeable difference is that people have a slight tendency to have front-back and up-down confusion [16] when non-individualised HRTFs are used.

3.2 Infrared LED tracking for detecting listener position and head movements

Since the distance perception has a direct influence of the externalisation and realism when listening to binaural audio, a positional tracking system is desirable. This can be done in various ways with e.g. facial recognition via a camera (in same style as Waves NX does described in 2.1.4). However, this does have its limitations as sufficient lighting should be present to track the face and get a proper frame rate. Several different proposed technologies have been been used in experimental setups; e.g. a head detection algorithm for tracking the listener's ears position in real-time using a laser scanner [11]. This method has proven to have very high accuracy (<= 15mm). However, this method requires expensive equipment and is not very convenient for commercial use. A different approach that does not include camera or sensor based tracking is an position estimation by acoustic signals only (e.g. voice or hand-clapping) [19]. It is achieved by the direction of arrival (DOA) from the acoustic source using two horizontally spaced microphones. This method can however be prone to issues as adverse effects by caused room reverberation can arise. Furthermore, this does not work as a real-time tracking system, but rather as a initial position calibration for the system.

A solution that eliminates these problems is the use of infrared (IR) tracking by having IR LEDs placed at the side of the listeners head and have an IR camera which only captures IR light from the LEDs for a more consistent and better refresh rate (up to 120hz depending on the camera being used). This is a well known method that is used for various applications such as head tracking for driving and flying simulation games. It is also a fairly simple and affordable way to create your own DIY head tracker.

Even though this approach is a better approach than the camera based facial recognition, it is still not completely optimal. Since it is still camera based, it will only be able to track what the camera can see. Therefore it can not track if you move outside of its range or rotate more than approximately 90 degrees in all directions. To compensate for this issue (at least with the limited rotation), an inertial measurement unit (IMU) can be implemented to track the head movement. In this way, the IR LEDs can be used for position and the IMU for head movement.

3.3 Design Requirements

Based on the two previous sections in chapter, investigating the state of the art and basing it on the problem statement, the following design requirements can be made:

1. Software implementation that uses a well known non-individualised HRTF dataset for binaural panning (Individualised HRTF will not be a focus as previous studies have found that sounds can still be perceived as externalised to the same degree whether or not the HRTFs are individualised).

2. Use infrared LED tracking instead of facial recognition. Improved refresh rate, accuracy and detected angles of rotation. It also enables positional tracking relative to the infrared camera that is being used.

3. Use fusion tracking that uses an IMU for tracking head movements and the IR LED to track the position.

Chapter 4

Design

In this chapter, the design and the implementation of the proposed system based on the state of the art and the design requirements will be outlined. The choices of the implemented technology both on the signal processing and the hardware part will be accounted for.

In order to create a system that simulates loudspeakers through headphones, there are two main focus areas which contain the DSP aspect that involves obtaining HRTFs, and applying that to the incoming sound for binaural playback through the headphones. The other is the head tracking that is mostly hardware based and is responsible for acquiring the positional and head movement data using a fusion of IR LEDs and an Intertial Measuring Unit (IMU) to obtain absolute positional tracking of the listeners head movement. This data is used for the HRTF algorithm to calculate the direction and distance from the sound source (loudspeakers) to the sound receiver (the listener).

4.1 Hardware

In this section the hardware will be outlined and accounted for. The choices of the chosen hardware is based on the research and what is most optimal for the use in this project.

4.1.1 Infrared tracking

It was chosen to do the tracking with the infrared solution as it seems to a reliable, affordable and stable solution for this project.

To track the IR LEDs, a customised clip was 3D printed to house the three LEDs. It is clip that was developed to easily apply to any headphone. The three LEDs (SFH485P, 880 nm) are in series connection wired to a USB cable to give the system 5V. The position of the LEDs are predefined by the recommendation of the

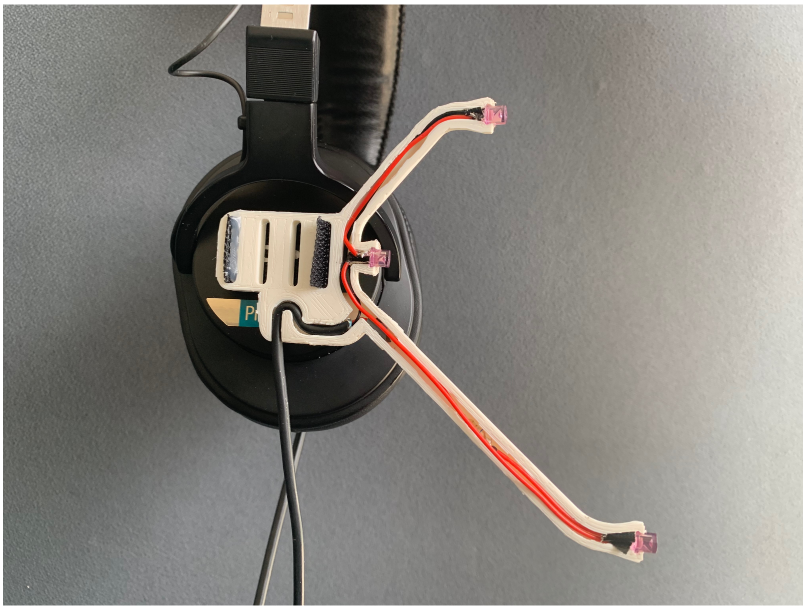


Figure 4.1: The custom 3D printed clip that houses three IR LEDs in a series connection connected to a USB cable which powers the circuit 5V. It is mounted onto the right side of Sony MDR-7506 headphones.

Pointtracker 1.1 software used in the open source tracking program OpenTrack to obtain the absolute position of the clip in 3D-space (XYZ). It is also possible to obtain the head movement orientation for its pitch yaw and roll using this technique. However, the maximum rotation that can be obtained is approximately 180° in all directions since the camera can not detect the LED when they are facing away from the camera. On figure 4.3 the Opentrack software captures the size and position of the LEDs relative to each other and the PointTracker 1.1 software thereby calculates the raw xyz and pitch, yaw and roll data. The yellow cross in the camera input shows the calibrated model center based on the three IR LEDs.

The camera that captures the position of the IR LEDs is a customised Sony PS3 Eye with its IR filter removed and a IR pass filter placed in front of the lens to only let IR light pass through. The camera operates with a resolution of 640x480 pixels and a frame rate of 60. The camera works best with little to no sunlight in the frame or any other IR light sources other than from the clip. It therefore works best indoors and without facing any windows where sunlight can hit the lens. It works perfectly in the dark and it always operates at the desired 60 frames per second.

4.1.2 Inertial Measurement Unit

A way to compensate for the limited tracking rotation is to use and additional sensor for head rotation measurement and let the IR LED clip account for the positional tracking only. An inertial measurement unit (IMU) can do the head movement sensing by measuring the pitch, yaw and roll when rotating the unit. However, since the sensor rely on a combination of different sensors in one (gyroscope, accelerometer and compass), it needs to be calibrated to give accurate measurements. During the design phase of the project small tests with an IMU (MPU-9250) connected to a Arduino Nano were carried out. However, the sensor shown too many inconsistencies in the measurements even after doing calibration and it was decided not to use an IMU for fusion sensing. The inconsistencies included drift of all angles over time which required the IMU to recalibrate every now and then. The system is intended to be used over longer periods of time, so the issues with drift of angles is not ideal. In order for maximum realism and the chance for perceived externalisation, the measurements need to be precise in order to give accurate audio feedback. It was therefore chosen to only use the IR LED clip for both positional and head tracking, but with the limitation of having about ± maximum 90° of rotation for pitch and yaw.

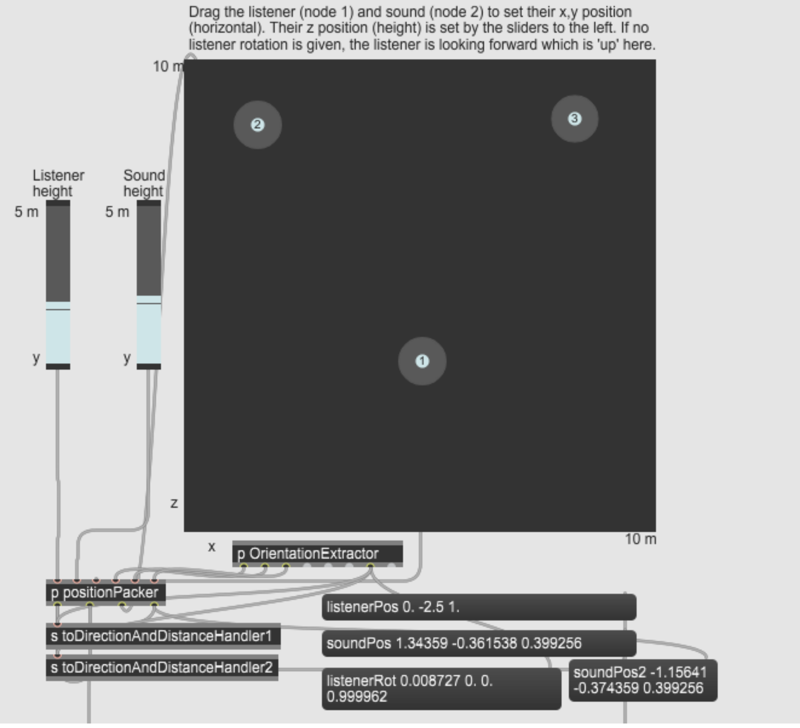
4.2 Software

In this section the software design and implementation behind the system will be described. A throughout description of the signal processing handled in Cycling '74 Max 8 [1] will be explained. The interconnection between the collection of head tracking data sent from OpenTrack to Max 8 will also be explained.

4.2.1 Max patch

The signal processing, HRTF calculation and relative distance sound emulation is developed in Cycling '74 Max 8 [1], based on the "FFT-based Binaural Panner" patch made by Jakob H. Andersen [3]. The main patch calculates the appropriate azimuth and elevation on the basis of a provided listener position. The two sound position objects are at a fixed position that are defined as the left and right speaker - in that way the patch can be set up as a virtual sound positioning system.

The patch contains a Java class that calculates the azimuth and elevation given the position of the listener based on the x, y, and z coordinates in relation to the sound positions coordinates. Furthermore, it also uses the listener rotation (unit quaternion) to calculate the azimuth and elevation. The distance from left to right ear based on the the rotation and position is also used to determine the interaural level, and time differences.



Drag the listener (node 1) and sound (node 2) to set their x,y position (horizontal). Their z position (height) is set by the sliders to the left. If no listener rotation is given, the listener is looking forward which is 'up' here.

listenerPos 0. -2.5 1.

soundPos 1.34359 -0.361538 0.399256

listenerRot 0.008727 0. 0. 0.999962

soundPos2 -1.15641 -0.374359 0.399256

Figure 4.2: The Max patch showing the position of the listener (1) and the left (2) and right (3) speaker on the 2D canvas.

Since the patch uses the CIPIC HRTF database, the patch needs to initialise a HRTF dataset from one of the participants from the database. Two matrix files containing the data to perform FFT are used for left and right and right channel. It runs with a signal vector of 1024 and 44100 Hz sample rate. The head related impulse response are converted to the frequency domain with a FFT size of 2048.

Inside the patch, it is possible to place the two sound sources (left and right speaker) on a 2D grid within a dimension that can be personally specified as seen on figure 4.2. Also, the listener position is also marked as a point on the canvas and updates in real-time and moves on the canvas accordingly. The ′OrientationExtractor′ object receives the listener′s head orientation (pitch, yaw and roll) and the position (x, y and z). This is feeded to the ′positionPacker′ object which packs messages the two sound sources and the listener and are seperately sent to the ′DirectionAndDistancehandler′ Java object for signal processing calculation for correct direction (azimuth and elevation) and distance from sound source to the listener.

Distance Emulation

To achieve realism for the system, a sense of distance from sound source to the listener, must be measured and calculated. One of the easiest way for humans to determine the distance to a sound source is the intensity of the sound - the further away the listener is to the source the more the intensity of the sound decreases.

Since the distance from the virtual loudspeaker to the listener is calculated, we must understand the relation between the intensity of a sound and how it propagates and reaches the listener at a certain distance. The radiation of sound loses power in proportion to the distance and loses about 3dB when doubling the distance [9]. This is given by the formula for Sound Intensity Level (SIL)

$$10\log_{10}(I) - 10\log_{10}(2I) = 10\log_{10}(1/2) = 3dB(loss) \tag{4.1}$$

The peak sound pressure of a sound wave is inversely proportional to the distance. Therefore, it decreases $1/r$ where r is the distance from the sound source. Given this information, the gain of the sound source at the specific position of the listener, both for the left and the right ear can be calculated in real-time. The calculation is made in the Java class 'DirectionAndDistanceHandler' that is responsible for all the appropriate distances and angles from the listener to the given sound emitter position. The distance (in centimeters) is updated directly from the measured IR positional tracking done in the Opentrack software.

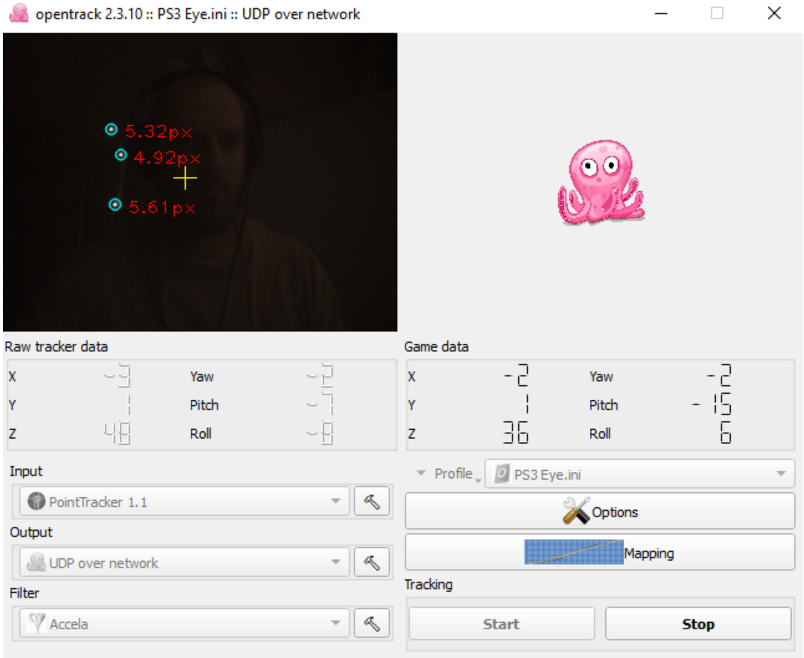
4.2.2 Opentrack integration and data flow

The Max patch needs the x,y and z and pitch, yaw and roll data in order to do the calculations for appropriate azimuth, elevation and distance from the sound source to the listener. The infrared LED tracking in Opentrack captures these position and head rotation data in centimeters and degrees. Fortunately, it is possible to send data from Opentrack to Max via an UDP protocol - Opentrack opens a port, sends compact packages via UDP and Max receives this data for further analysis.

Max does not natively have an object that can receive these packages, so a third party tool called Sadam library that can handle binary streaming - it receives the data from Opentrack and converts it to bytes. The data is then split up to each of its own and the head movement data is converted into unit quaternion using the euler2quat object. The combined quaternion or xyz coordinates with a prepended 'listenerRot' or 'listenerPos' are sent as a message to the 'DirectionAndDistanceHandler' Java object.

4.3 Aim of study

The choice of design and implementation has its purpose to create a system that is easy accessible and can simulate a loudspeaker set through a pair of headphones. The next step is to make an experiment to test out if an user of the system can perceive it as the sound is coming from "outside of the head" - the music is externalised.



opentrack 2.3.10 :: PS3 Eye.ini :: UDP over network

Raw tracker data

X	-3	Yaw	-2
Y	1	Pitch	-7
Z	48	Roll	-8

Game data

X	-2	Yaw	-2
Y	1	Pitch	-15
Z	36	Roll	6

Input: PointTracker 1.1

Output: UDP over network

Filter: Accela

Profile: PS3 Eye.ini

Figure 4.3: The OpenTrack 2.3.1 software detecting the three IR LEDs (the yellow cross indicates the model position after calibrating) and by using the Pointracker 1.1, it calculates the raw tracker data with head movement and positional data.

Furthermore, it will also be to test if the IR LED tracking that has 6 DOF (both head movement and position) does improve the externalisation. The focus of presented research ([16], [7]) focuses mainly on the head movement only, hence it would be interesting to see if the distance implementation has an enhanced effect on the externalisation aspect.

It has been chosen to use non-individualised HRTFs mainly because personalised HRTFs would be too cumbersome to measure for each participant and also because this would contravene the objective of the system to be easy accessible. However, it must be investigated if the use of non-individualised HRTFs has a negative affect on the localisation cues and externalisation. There will not be a comparison between individualised and non-individualised HRTFs, but an investigation for the listeners if the sound on this system sounds realistic or if it is disorienting because of the HRTFs not being personalised.

Given these points of interests, a hypothesis can be made which will be the focus of the experiment.

H0 The addition of positional tracking compared to only head movement tracking will increase the reported externalisation of the subjects.

Chapter 5

Experimental Setup

The experiment is inspired by the test conducted by Hendrickx *et al.* [16], which they reproduced Brimijoin *et al.* [7] experiment. The focus of their rendition of the experiment was to see if large head movements (±90°) had a significant improvement on externalisation. They also wanted to redo the experiment to see if subjects could determine the degree of externalisation after they had stopped moving his/her head. Hendrickx *et al.* also found a lack of detailed data to support the claims they made, as subject could more or less move freely. They wanted the movement to follow a specific protocol to make sure that subjects had the same movement and thus can more confidently reject or accept their hypothesis. Also, they wanted to see if the use of non-individualised HRTFs could be applied as it would represent a more generalisable display scenario. Lastly the stimulus being used was longer than the previous one being used (8s instead of 2-3s). This was to make sure the subjects had enough time to determine the degree of externalisation and make large head movements as well.

They presented three hypothesis with the focus that large head movements would improve externalisation when the head tracker is active and a collapse of externalisation will happen when the head tracker is inactive. They found in their experiment that indeed head movements coupled with head tracking led to a substantial improvements of externalisation for most subjects. In the present study, it will be assumed that this condition is true, but the additional positional tracking coupled with head tracking will even further improve externalisation compared to only head tracking. The choice to replicate the experiment is to foremost have a valid test and to also compare it Hendrickx *et al.*'s findings. More or less the same procedure and protocol will be used in respect of the head movements and post condition externalisation questions.

5.1 Stimulus

In Hendrickx *et al.* [16] they claimed that the 2-3s stimulus used in the experiment conducted by Brimijoin *et al.* [7] was too short for participants to determine the degree of externalisation and used a 8s excerpt of male speech instead. Even though this stimulus is longer the author of the present study still found the stimulus to be too short for subjects to make large enough head movements, get familiar with the sound and to determine the degree of externalisation. This is also due to the participants should also do more movement and not only head rotation. Because of this a 30s stimulus was chosen for the test.

It was also chosen to use music as the stimulus instead of speech. It was better suited to the overall problem statement of simulating a loudspeaker setup having two speaker for a left and right channel. The music is the first 30 seconds of the Paul McCartney's "Fool on the hill". It consists of piano, drums, guitar, flute and singing, which tries to cover most of the frequency spectrum and consists of transient and sustained sounds.

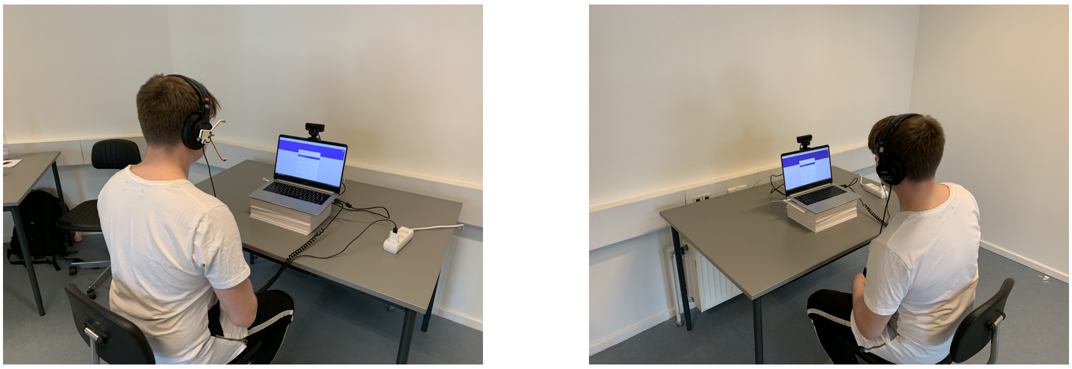
5.2 Location and Experimental Protocol

The experiment took place at Aalborg University in Copenhagen in a small room to ensure no disruption and environmental noise. An introductory questionnaire was presented with demographic questions and their experience with music production and familiarity with HRTF and binaural audio. The subjects had to follow 4 different head movement protocols that was explained by the test conductor. The four conditions are inspired by the ones used in Hendrickx *et al.* experiment, but with slight modification with added positional movement:

- **NH** : Head orientation (± 90° left and right), no head tracking.
- **NP** : Head orientation (± 90° left and right) + position (back/forth, side/side), no head tracking.
- **WH** : Head orientation (± 90° left and right), with head tracking.
- **WP** : Head orientation (± 90° left and right) + position (back/forth, side/side), with head tracking.

When the music started the subject performed the head and/or position movement until the 30 seconds of stimulus was over. They could repeat the movement routine if they wanted to. All subjects did the same movements to ensure that everyone received the same cues and to make a more valid comparison.

After each condition, the subject had to report their degree of externalisation of the music from a scale from 0-5, where 0 is "The source is at the center of my head"



(a) **(b)**

Figure 5.1: The test setup while a subject is performing one of the conditions. On (a) the IR LED clip can be seen on the right side of the headphones. The PS3 eye camera placed on top of the laptop tracks the position of the clip in real-time.

and 5 is "The source is remote and externalised. This was to ensure that the subject would report the after effects of the externalisation just after each condition. This is the same scale and questions used by Hendrickx *et al.* [16] in their experiment.

Lastly, the subject was presented with a customised System Usability Questionnaire (SUS) consisting of 8 questions with a focus of the system's responsiveness, audio quality and feedback.

Chapter 6

Results

A total of 20 subjects participated in the experiment (16 men and 4 women, aged 20-26 years). Out of the 20, 11 of the subjects compose music and does their mix on their computer/laptop. Out the the 11 subjects most of them (7 out of 11) usually use headphones while mixing while the rest usually use loudspeakers. This is to support the claim that a fair amount of people and the majority of music producers (mostly as a hobby) often use headphones as their main source for audio feedback.

Nine of the of the twenty participants were familiar with the terms and "binaural hearing" and/or "HRTF".

6.1 Difference with head movement and added position tracking

With the hypothesis "the addition of positional tracking compared to only head movement tracking will improve externalisation." it will be interesting to compare condition **WH** with **WP**.

The experimental protocol is a repeated measure for the same group of people for condition **WH** and **WP**, where the difference is the movement protocol - **WH** only consisting of head movement and **WP** consisting of head movement + positional movement. For this purpose a one-tailed paired t-test was used to determine if there is a significant difference between the two groups. Hypothetically, the **WP** condition should have a significantly higher externalisation score than the **WH** condition. The mean scores for condition **WH** and **WP** were, 3.05 ($STD = 1.234$) and 3.30 ($STD = 0.979$) respectively. Given the statistical test there was no significant difference found between head movement versus head movement + positional movement in respect of externalisation ($p = 0.15$).

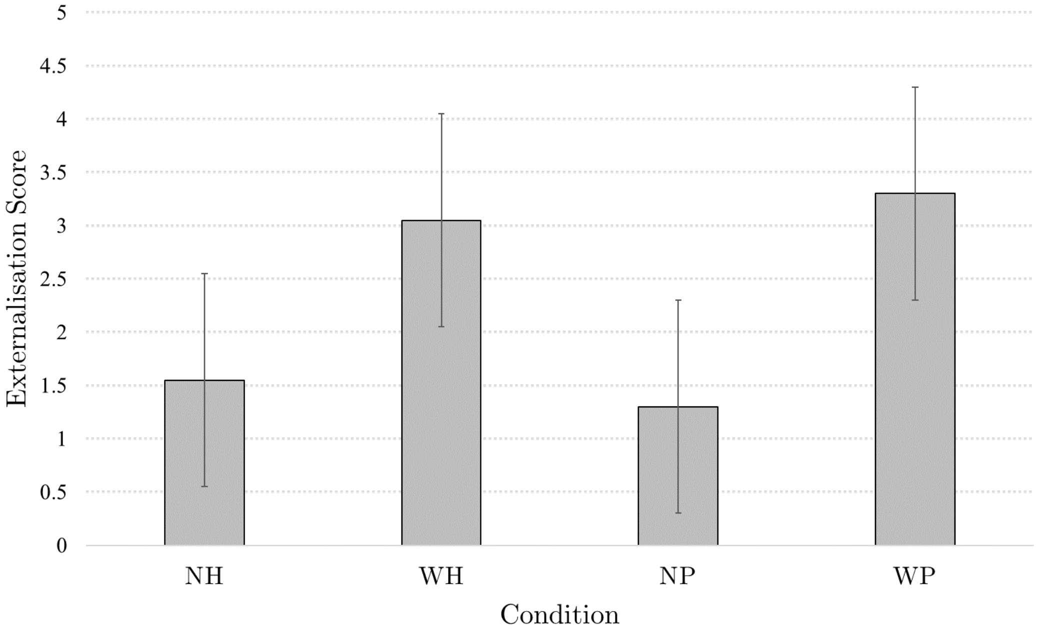


Figure 6.1: Mean externalisation scores for all four conditions. An error bar indicating the standard deviation ($NH = 1.09, WH = 1.23, NP = 1.17, WP = 0.98$) for each condition is also shown. **NH**: Head orientation, no head tracking. **WH**: Head orientation, with head tracking. **NP**: Head orientation + position, no head tracking. **WP**: Head orientation + position, with head tracking.

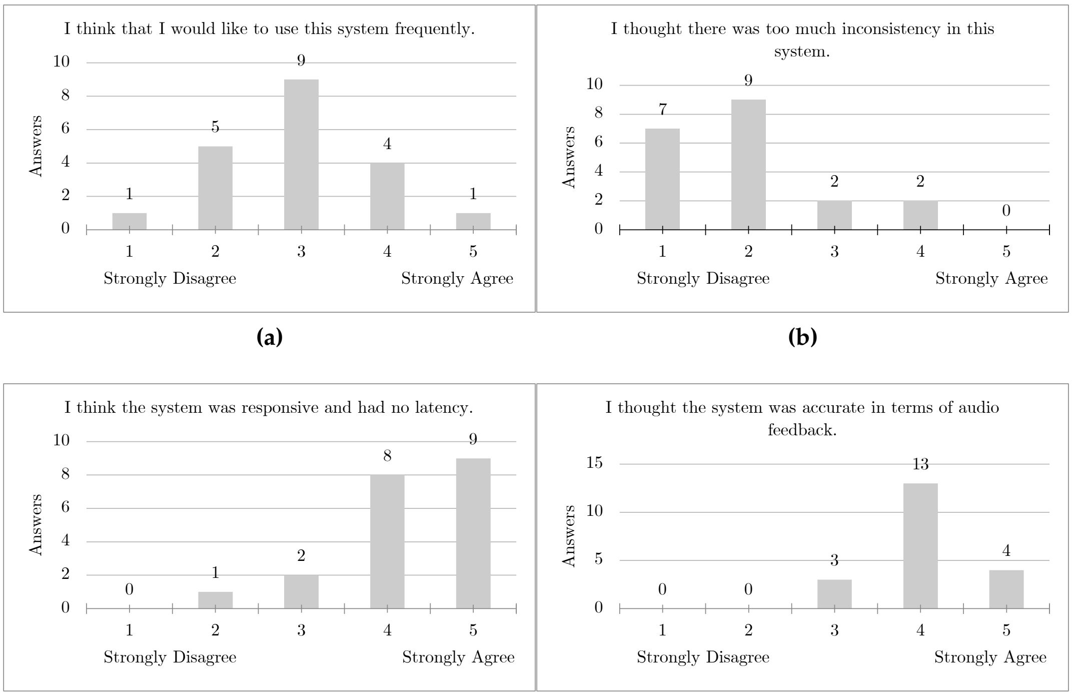
6.2 Data logging of movement

During the test, the movement data was recorded for each condition[1]. This includes the pitch, yaw and roll (head movement) and the x, y and z coordinates (positional movement). This was done in OpenTrack software that saved the data to a CSV file while doing head movement tracking. During the test, the movement data sent to the Max patch was the filtered data by using the Acella filter made by Stanislaw Halik [15].

No further insight was made with this data, but more as a confirmation that the tracking was undergoing and the subject would consequently hear the correct binaural audio based on the tracked data. There was found no inconsistencies reviewing the logged tracking data. Inconsistencies would be stuck tracking (the infrared camera unable to track the IR LEDs and consequently being stuck on the last recorded movement) or incorrect head movement logging (the software detecting more IR LED spots than the desired 3, consequently making false calculations).

Nine of the participants were familiar with the terms "binaural hearing" and/or "HRTF". This group rated the degree of externalisation a mean scores of the **WH**

[1] All of the logged data can be downloaded and review from here `https://bit.ly/2Ma1bOB`



(a) **(b)**

(c) **(d)**

Figure 6.2: Diagrams showing the SUS responses on four of the eight statements presented to the subjects.

and **WP** 3.0 and 3.7, respectively. A T-test also found a significant difference between these two groups ($p = 0.01$). However, there was not a significant difference for the group that were not familiar with the terms ($p = 0.4$) given the mean scores 3.1 and 3.0 for **WH** and **WP**, respectively. This could indicate that the subjects that were familiar binaural audio and might have tried it before, could notice the difference between the two conditions and hence rate a higher degree of externalisation.

6.3 Difference between conditions with and without head tracking

Given the mean values of reported externalisation as seen on figure 6.1, it can be seen that there is a significant difference in the scores of the conditions without tracking (**NH** and **NP**) with the conditions with tracking (**WH** and **WP**) with a p value of 0.0000001. This was done to partly confirm that using headphones and making head movements with no tracking will keep the sounds internalised and appear to be inside the skull. It was also make a reasonable comparison with the condition that included head movement tracking.

6.4 System Usability Scores

The subjects answered 8 SUS statements from a scale from 1-5 with 1 labelled "Strongly Disagree" and 5 labelled "Strongly Agree". On figure 6.2 4 of the SUS results can be seen[2].

The System usability questions after the test showed that they generally felt confident using the system, thought it was responsive and had no latency and that the audio feedback was accurate in respect to their movements. The subjects were neither agreeing or disagreeing with the statement that they wanted to use the system frequently for use to when producing music. Some mentioned that they could imagine them using it watching a movie or while playing a video game and thought it might not be useful for music production. During the experiment, some subjects might have experienced that the tracking was inconsistent, if they for example rotated too much and the camera could not detect the IR LEDs or if background IR light (from e.g. sunlight) was detected and falsely measured as part of the LED clip. However, none of the participants reported these inconsistencies and this was also not observed through the logged head movement data.

6.5 Summary

The results of the present study can be summarised as follows:

- There was not found a significant difference between condition **WH** and **WP** - the addition of positional tracking does not certainly improve externalisation in this experiment.
- In the conditions **WH** and **WP** where subjects were familiar with the terms "HRTF" and "binaural hearing", a significant difference were found between the two conditions compared to the group that were not familiar with the terms. This could indicate that subjects were aware of the addition of position tracking which affected the music and hence improved the externalisation.
- The subjects generally rated the SUS statements positively and they thought the system was responsive, had no latency and was accurate in terms of audio feedback in respect to their movements.

[2] A full review of all results including demographics, reported externalisation and SUS scores can be found on `https://bit.ly/2I0jUXG`

Chapter 7

Discussion

7.1 Comparison with studies

Hendrickx *et al.* [16] made their experiment because they found that in previous studies results showed that head movements did not enhance externalisation and in general the role of head movements in the phenomenon of externalisation remains unclear. They found a lack of sufficient subject and quantative data to fully conclude on this research question. Therefore, they replicated the experiment performed by Brimijoin *et al.* [7], but with some adjustments to make a more valid test. This included amongst others, a longer stimulus, a more controlled and strict protocol for head movements and make sure that the externalisation is reported after the subject has moved his/her head (aftereffects).

In the present study, it was chosen to use the experiment of Hendrickx et al. with slight modifications to make it suit well to the hypothesis made to in respect to the current research question. Furthermore, it was chosen to have a substantially longer stimulus (from 8s to 30s) to make sure the subjects had enough time to make the movements that was intended. For some subjects they repeated the movements, others did it once, but everyone made at least a full cycle of the movements demonstrated by the test conductor. It was also chosen to choose the same six-point scale to report externalisation. This was preferred as some previous studies asked the degree of externalisation during the stimulus and others had a binary answer to which the subject could only answer that the stimulus was inside the head (internalised) or outside of the head (externalised).

Hendrickx et al. found that head movements coupled with head tracking did substantially enhance externalisation, compared to a situation where the listener does not move the head. They also found that the head movements coupled with head tracking enhance externalisation to an even further extent, compared to the situation where the listener moves his/her head but without head tracking. In the present study, the premise was to found out if the addition of positional tracking did enhance externalisation to an even further extent, compared to the situation with head movements coupled to head tracking.

It was indeed found that in the condition with head movement coupled with head tracking had a higher externalisation score **NH** (m = 3.05) compared with the one without head tracking **WH** (m = 1.55) and the two groups were significantly different (p = 0.0003). The same is applicable with the condition with head movements + position coupled with head and positional tracking **NP** (m = 1.30) compared with the same condition but without tracking **WP** (m = 3.30, p = 0.00006). The present study can therefore confirm that the externalisation collapses once the subjects does head movement and/or positional movement but without tracking enabled. However, it cannot be confirmed that the addition of positional movement and tracking does significantly increase the degree of externalisation.

Hendrickx et al. suggests that head movements may need to be sufficiently large in order to have an impact on externalisation [16]. This might be the reason because condition **WP** did not have large enough movement from side/side and front/back. Since the limitation of camera-based tracking, the subject could only be within a certain frame. The maximum side/side movement was preferred and it span from 20-30 cm to the side from the center position. The back and forth motion also had its limitations because if the subject came too close, the camera might not detect the three infrared lights and would give incorrect tracking. Also, if the subjects moved too far away, it would be harder to detect the lights especially if head movements to the sides were made. However, the test conductor made sure that the subjects moved approximately 50cm both back and forth.

7.2 Different Conditional Testing

In the present experiment there could be a condition with positional movement, but only with head movement tracking versus positional movement and positional tracking. In that way it would be easier to compare if positional tracking improved externalisation.

The current test setup is designed to compare the direct difference of the addition of positional movement has an improvement to externalisation compared to only head movement. For condition **WH** and **WP**, the head tracking and positional tracking is active. The effects of positional tracking which emulates the distance (gain of the sound relative to the distance from sound source to subject) is still active in the **WH** condition. This might be the reason that there was not found a significant difference between the two conditions - the conditions were too close in terms of audio feedback. Alternatively, it would be interesting to have another condition similar to **WP** but which only has head movement tracking enabled – disabling the positional tracking. In that way it would be possible to compare this condition with **WP** and see if the positional tracking has a direct affect on externalisation compared to only head movement tracking.

7.3 Visual cues to improve externalisation

During the experiment the subjects were facing a laptop without having anything that visually represented the sound source and its location. The subjects were not informed of the actual position of the sound source relative to their position and how far it was away from them. It might be a difficult task for them to determine the degree of externalisation without knowing the approximate distance to the source and without having a visual cue that indicates its position. However, this was not the focus of the experiment and was also not part of how Hendrickx et al. set up their experiment and what the current experiment is based on. Nonetheless, it is an interesting point of interest that might have an effect of the whole experience for the listener.

Brimijoin et al. [7] discusses this issue in their research as well and draws a connection between the presence of a visual target and the greater degree of externalisation. In their experiment, they had visible loudspeakers that propagated their sounds, but was maybe less noticeable as it was a loudspeaker ring surrounding the participant, which created many possible visible targets.

The presence of a visual target and draw the perceived location of the sound towards it, is called the ventriloquist effect. It would be interesting to implement visual targets representing loudspeakers for the current experiment and see if this improves the externalisation for the subjects.

7.4 Influence of non-individualised HRTFs

In the present experiment the HRTF data from the third subject in the CIPIC HRTF database was used. This was not chosen for a particular reason, but rather because this was the default one originally used in Hougaard Andersen's FFT-based binaural panner Max patch [3]. It was chosen to use non-individualised HRTFs from the *Design Requirements* since several studies did not report a significant difference in externalisation between individualised and non-individualised binaural synthesis with speech stimuli. This includes studies from Møller *et al.* [20] and Begault *et al.* [4]. Recently, Geronazzo *et al.* [12] found with their anthropometry based mismatch function that there exists a non-individualised HRTF set that allows a listener to have an equally accurate elevation localisation than with individual HRTFs. They also used the CIPIC HRTF database for their research as their non-individualised HRTF sets alongside with the HRTF database from the Acoustics Research Institute (ARI). Not only did the listener have a accurate localisation, but it also enhanced the externalisation and the up/down confusion rates.

These studies shows that individualised HRTFs is not necessary (although ideal) as non-individualised HRTFs are just as accurate and some cases equal to the individualised ones. However, HRTFs that are selected based on the anthropometrc data (distance from ear to ear, size of pinna etc.) of the listener is recommended as a preliminary study by Geronazzo *et al.* [13] showed. They found that selecting the HRTF based on mismatch function that relies on the anthropometric data of the listener increased the average performances of 17% for elevation accuracy compared to the use of a generic HRTF with anthropometric data. It also significantly increased externalisation and up/down confusion rates. This kind of customisation of HRTFs is already seen in some of the state of the art explained in the *Related Work* chapter. This includes the *Waves Nx* software that allows the user to customise their HRTF with the "head modelling" that uses the measured circumfrence and inter-aural arc to calculate delays, filters and the levels of the sound for each ear. The other example, *Ossic X*, that does not need the user to specify the anthropometric features themselves, but instead the headphone automatically measures it when the headphone is mounted on the head.

7.5 Implement reverberation to better simulate the current room

It is known that reported externalisation is strongly linked with the amount of reverberation to the stimuli. In the experiment performed by Begault *et al.* [4] they found that an anechoic stimuli participants made an externalised judgement of 40% compared to 79% of a reverberant condition (the subjects could score the degree of externalisation between 0-100%). They also found that there was no significant difference between a early-reflection and a full-reverberation condition. This means that an externalised stimuli can be simulated using a minimal representation of the acoustic environment the subject is in.

It was also discussed in the introduction chapter that reverberation would be a improvement to the perceived realism of the stimuli. However, the term "realism" in this sense can be a wrong term to use. Begault *et al.* [4] asked their participants to rate the perceived realism and they did not find significant effects and the lack of variability might suggest that the participants did not differentiate among the conditions based on the perceived realism, or that they simply did not have a common understanding of what "realism" meant. However, the term externalisation could perhaps still be used as it seems that participants both in Hendrickx *et al.* and the current experiment have a good understanding with the term "externalisation" given the six-point scale with explanation they were provided to answer after each condition.

It would be interesting to see if the addition of a reverberated stimuli that is either full-reverberated or with early reflection can increase the degree of externalisation of the conditions even further. This would also include a room simulation that changes the early reflections and amount of reverb based on the distance from the subject to the sound source.

Chapter 8

Conclusion

This paper has presented the design and implementation of a positional IR tracking based system that simulates a loudspeaker setup using non-individualised HRTFs. There has been an investigation of the state of the art within binaural and spatialised audio both for software and hardware. Many of the software solutions presented has sophisticated binaural solutions with virtual acoustics and customised HRTF features, but most of them only focuses on the azimuth and elevation and not the position of either the sound source or the listener.

Based on these investigations, design requirements were made with the intention to make a system that can do positional tracking to simulate a loudspeaker setup through headphones using non-individualised HRTFs. The implementation of the system includes positional and head movement tracking using 3 point infrared LED clip attached on the side of a pair of headphones and a camera to capture head movement and positional distance sensing relative to the camera.

An experiment including 20 subjects tried out the system while performing different conditions with different movements with or without tracking. The subjects rated the degree of externalisation of the stimuli after each condition.

The results showed that there was no significant difference in the degree of externalisation between the condition only having head movement and the other that had the addition of positional movement and tracking. However, there was found a significant difference between the group of subjects that were familiar with binaural hearing in the conditions with only head movements and additional positional tracking.

These findings were discussed and compared with previous studies that had similar experimental setup. Several points of interests were presented that could improve or might enhance the externalisation to the already proposed system. This includes implementing virtual acoustics i.e. reverberation and including anthropometric data matching for subjects for determining a suitable HRTF set for the listener.

Bibliography

[1] Cycling '74. *Cycling '74 Max 8*. 2019. URL: `https://cycling74.com/`.

[2] *The CIPIC HRTF database*. English. 2001, pp. 99–102. ISBN: 0-7803-7126-7. DOI: `10.1109/ASPAA.2001.969552`.

[3] Jakob Hougaard Andersen. *FFT-based binaural panner*. 2019. URL: `https://cycling74.com/tools/fft-based-binaural-panner`.

[4] Durand Begault and Elizabeth Wenzel. "Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source". In: Oct. 2001, pp. 904–916.

[5] Laurent Betbeder. *Near-field 3D Audio Explained | Oculus*. 2017. URL: `https://developer.oculus.com/blog/near-field-3d-audio-explained/`.

[6] Jens Blauert. *The Technology of Binaural Listening / edited by Jens Blauert*. eng. Elektronisk udgave. Modern Acoustics and Signal Processing. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. ISBN: 9783642377624.

[7] W. Owen Brimijoin, Alan W. Boyd, and Michael A. Akeroyd. "The Contribution of Head Movement to the Externalization and Internalization of Sounds". English. In: 8 (2013), e83068–. ISSN: 1932-6203. DOI: `10.1371/journal.pone.0083068`.

[8] dearVR. *dearVR MUSIC - Dear Reality*. 2019. URL: `https://www.dearvr.com/products/dearvr-music`.

[9] Andy Farnell and Farnell. *Designing Sound*. English. 2010. ISBN: 978-0-262-01441-0.

[10] Kim F Fluitt, Timothy Mermagen, and Tomasz Letowski. "Auditory perception in open field: Distance estimation". In: (2013).

[11] Panayiotis G Georgiou et al. "Immersive sound rendering using laser-based tracking". In: *Audio Engineering Society Convention 109*. Audio Engineering Society. 2000.

[12] Michele Geronazzo, Simone Spagnol, and Federico Avanzini. "Do We Need Individual Head-Related Transfer Functions for Vertical Localization? The Case Study of a Spectral Notch Distance Metric". English. In: 26 (2018), pp. 1247–1260. ISSN: 2329-9290.

[13] *Enhancing vertical localization with image-guided selection of non-individual head-related transfer functions*. English. 2014, pp. 4463–4467. ISBN: 978-1-4799-2893-4.

[14] G.R.A.S. *GRAS Sound & Vibration*. 2019. URL: `http://www.kemar.us/`.

[15] Stanislaw Halik. *Accela in opentrack 2.3*. 2019. URL: `https://github.com/opentrack/opentrack/wiki/Accela-in-opentrack-2.3`.

[16] Etienne Hendrickx et al. "Influence of head tracking on the externalization of speech stimuli for non-individualized binaural synthesis". English. In: 141 (2017), p. 2011. ISSN: 1520-8524. DOI: `10.1121/1.4978612`.

[17] HE Jianjun, Ee Leng Tan, Woon-Seng Gan, et al. "Natural sound rendering for headphones: integration of signal processing techniques". In: *IEEE Signal Processing Magazine* 32.2 (2015), pp. 100–113.

[18] Kickstarter. *OSSIC X: The first 3D audio headphones calibrated to you by OSSIC*. 2019. URL: `https://www.kickstarter.com/projects/248983394/ossic-x-the-first-3d-audio-headphones-calibrated-t`.

[19] Ki Seung Lee and Seok Pil Lee. "A real-time audio system for adjusting the sweet spot to the listener's position". French. In: *IEEE Transactions on Consumer Electronics* 56 (2010), pp. 835–843. ISSN: 0098-3063. DOI: `10.1109/TCE.2010.5506009`.

[20] Henrik Møller et al. "Binaural Technique: Do We Need Individual Recordings?" In: *J. Audio Eng. Soc* 44.6 (1996), pp. 451–469. URL: `http://www.aes.org/e-lib/browse.cfm?elib=7897`.

[21] Oculus. *Localization and the Human Auditory System*. 2019. URL: `https://developer.oculus.com/documentation/audiosdk/latest/concepts/audio-intro-localization`.

[22] Sennheiser. *Ambeo Orbit*. 2019. URL: `htpps://sennheiser.com/ambeo-blueprints-downloads`.

[23] Sennheiser. *Sennheiser AMBEO Smart Headset - Headset til mobile binaurale optagelser*. URL: `https://da-dk.sennheiser.com/finalstop`.

[24] Kasper Skov. *Binaural spatialization in Ableton*. 2019. URL: `http://kasperskov.dk/projects_binaural_jit.html`.

[25] Waves. *Nx - 3D Audio on Any Headpones | Waves*. 2019. URL: `https://www.waves.com/nx`.

[26] Elizabeth M. Wenzel et al. “Localization using nonindividualized head?related transfer functions”. English. In: 94 (1993), pp. 111–123. ISSN: 0001-4966. DOI: `10.1121/1.407089`.

[27] Tomasz Woźniak. *HRTF | Code and Sound*. 2015. URL: `https://codeandsound.wordpress.com/tag/hrtf/`.