Influence of Elevation on Localization Performance in Mobile Audio Augmented Reality Systems

Bachelor Thesis
The present work was submitted to the
Chair of Computer Science 10
Prof. Dr. Jan Borchers
Computer Science Department
RWTH Aachen University

by
Jayan Jevanesan

Thesis advisor:
Prof. Dr. Jan Borchers

Second examiner:
Prof. Dr. Torsten Kuhlen

Registration date: 03.08.2015
Submission date: 08.10.2015

I hereby declare that I have created this work completely on my own and used no other sources or tools than the ones listed, and that I have marked any citations accordingly.

Hiermit versichere ich, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie Zitate kenntlich gemacht habe.

Aachen, October 2015
Jayan Jevanesan

Contents

List of Figures

List of Tables

Abstract

Mobile Audio Augmented Reality Systems are used to create virtual auditory spaces by externalizing (virtualizing) the sound which can then be presented over headphones. Using binaural audio rendering technique, the audio is spatialized and gives the listener the impression that the sound is coming from a fixed position in the physical space. Rendering the audio in real-time and tracking the movements with a motion-tracker allows users to navigate in the virtual audio space. Most of the implementation and research done on audio augmented reality only involve rendering the horizontal plane and orientation of the listeners. In this thesis we are going to examine how the simulation of the median plane in virtual audio environments influences the listener's performance on localizing virtual sound sources and how this might affect the overall perception of the virtual audio space.
To acquire these results two studies have been conducted. The first study gives us information about the minimal audible angle in the median and horizontal plane of the rendering algorithm we used. The results of the second study show us how additionally simulating the elevation impacts the localization of sound sources in the horizontal plane.
The results of this thesis provide information on whether the integration of the vertical axis into a virtual audio space would improve the localization performance and the overall perception of the virtual audio space.

Überblick

Mobile Audio Augmented Reality Systeme werden verwendet um virtuelle akustische Räume durch Virtualisierung des Klangs zu erschaffen, welche dann über Kopfhörer ausgegeben werden können. Durch binaurales Audio-Rendering kann man Klänge eine räumliche Eigenschaft verleihen um dem Zuhörer den Eindruck zu vermitteln dass sich die Tonquellen verstreut auf festen Positionen im selben Raum wie er selbst befinden. Damit man sich in diesem virtuellen Umfeld bewegen kann, wird der Ton in Echtzeit gerendert während ein Motion-Tracker die Bewegungen des Benutzers erfasst. Die meisten Systeme und Forschungen im Bereich Audio Augmented Reality berücksichtigen nur die horizontale Ebene und Orientierung. In dieser Arbeit werden wir den Einfluss der Simulation der vertikalen Ebene auf die Leistung der Lokalisierung und Wahrnehmung von Soundquellen in einem virtuellen Umfeld untersuchen.
Dazu haben wir 2 Studien durchgeführt. Die erste Studie gibt uns Informationen über den minimalen hörbaren Winkel in der horizontalen und vertikalen Ebene für den von uns benutzten Rendering-Algorithmus. Die zweite Studie gibt Informationen über den Einfluss der zusätzlichen vertikalen Ebene auf die Lokalisierung von Soundquellen in der horizontalen Ebene.
Die Ergebnisse dieser Studie geben uns Auskunft darüber ob die Einführung der vertikalen Ebene in einem virtuellen Audioumfeld die Lokalisierung von Soundquellen und das Hörerlebnis verbessert.

Resumé

Les systèmes mobiles de réalité augmentée audio sont utilisés pour créer des espaces virtuels acoustiques en virtualisant le son qui est transmis via un casque. Grâce à la synthèse binaurale, le son peut avoir un caractéristique spatiale et donne l'impression à l'auditeur que les sources sonores sont dispersées dans des positions fixes dans l'espace physique. Le son peut être synthétisé en temps réel en captant les mouvements de l'utilisateur et ainsi lui permet de se déplacer dans cet espace virtuel. La plupart des systèmes et études de ce domaine considèrent que le plan et l'orientation horizontale. Dans cette thèse nous allons étudier l'influence de la simulation d'un plan verticale sur la performance de la localisation des sources sonores dans un espace virtuel acoustique. Pour déterminer ceçi nous avons mené deux études. La première nous donne des informations sur le seuil de déplacement angulaire dans le plan horizontal et vertical, alors que les résultats de la deuxième étude montre comment la simulation supplémentaire du plan vertical affecte la localization de sources sonores dans le plan horizontal et influence la perception de l'espace virtuel acoustique.
Les résultats de cette étude nous fournissent des informations si l'introduction du plan vertical dans un espace virtuel acoustique permettrait d'améliorer la performance de la localisation auditive et la perception de l'espace virtuel.

Acknowledgements

First of all I would like to thank Florian Heller for supervising my thesis and for his consistent help. I also would like to thank Pascal and Roman from KLANG:technologies Gmbh for their equipment and for helping us setup the experiment. At last I want to thank all of the users who participated in all of the studies.

Conventions

Throughout this thesis we use the following conventions.

Text conventions

Definitions of technical terms or short excursus are set off in coloured boxes.

EXCURSUS:
Excursus are detailed discussions of a particular point in a book, usually in an appendix, or digressions in a written text.

Definition:
Excursus

Source code and implementation symbols are written in typewriter-style text.

```
myClass
```

The whole thesis is written in American English.

Download links are set off in coloured boxes.

File: myFile[a]

[a] http://hci.rwth-aachen.de/public/folder/file_number.file

Chapter 1

Introduction

With the emergence of smartphones, tablets and other mobile devices with high processing power, mobile virtual reality systems have become an important part of many interactive media systems [Sander et al., 2012]. Audio augmented reality systems create a complete audible virtual environment in which the users can move around e.g. museums Heller [2014], Terrenghi and Zimmermann [2004] or even in fighter aircrafts [Jan Abildgaard Pedersen, 2005]. By modifying the audio which is emitted through headphones in relation to the user's position and orientation, the sound is perceived in a fixed position in the physical space.

Most of the current audio augmented reality systems only feature the horizontal plane, i.e., navigating around in two dimensions by rendering the audio based on only the location and landscape orientation of the user, often equipped with a GPS system and a compass. This might be sufficient for some use cases, however there is a lack of realism since the real world is made of three dimensions. Current mobile devices have the capability to run more complex audio rendering algorithms with higher fidelity while modern sensors can measure head movements very accurately. This enables new possibilities such as tracking and rendering elevation. Adding the third dimension, i.e., elevation into virtual audio environments might improve the localization of sources in the horizontal plane, thus increase the orientation performance and enhance the

Three dimensions to increase realism.

experience itself. The improved localization of sources could even lead to a more flexible placement of sources in the horizontal plane, for example two unmovable visual sources from which virtual sounds are emitted, could be placed too close to each other so that it is not possible to separate the virtual sound sources bound to these visual sources. Thus additionally simulating the elevation might help discerning the two sources.

Simulating the horizontal and vertical plane means that for the audio to be rendered accordingly, it has to be consistent with the movements of the user. The direction of the sound is localized by two main factors, the interaural time difference (ITD) and the interaural level difference (ILD).

Definition: *ITD and ILD*

ITD AND ILD:
The two main cues of sound source localization are the interaural time difference (ITD) and the interaural level difference (ILD), which are caused by the wave propagation time difference and the shadowing effect by the head, respectively [Pulkki, 1997].
ITD is the difference in arrival times of a signal reaching both ears. ILD is the difference in amplitude (loudness) of a signal reaching both ears

Higher precision achieved with HRTFs.

These can be emulated by rendering the audio appropriately, the audio environment can then be created through headphones so that sound sources appear at fixed positions in the physical space. Amplitude-panning [Pulkki and Karjalainen, 2001] for example is one approach to do so, this simple technique involves changing the loudness of the stereo channels such that the audio is perceived in the desired direction. However increased spatial resolution and simulating elevation is attained when rendering is done using the head-related transfer function (HRTF) in the time domain called head-related impulse response (HRIR).

HEAD-RELATED TRANSFER FUNCTION (HRTF):
a complex-valued free-field transfer function from a sound source in a certain direction to the eardrum [Bronkhorst, 1995].

Definition: *Head-Related Transfer Function (HRTF)*

The HRTF is usually measured in an anechoic room with in-ear microphones and is unique for each person. Localization of virtual sources with the use of individualized HRTFs is almost as accurate as with real sources [Bronkhorst, 1995] but since it is too effortful to determine the HRTF for each individual, general HRTFs are usually used. Using general HRTFs, i.e., the HRTF of another human individual or dummy-head [Zhang et al., 2009], can however result in inaccurate localization [Bronkhorst, 1995] with confusion errors (front-back and up-down) being problematic (cf. Chapter 2 "Related work"). These can decrease with increasing experience as do the localization errors, since there is a "learning-effect", i.e., the auditory system of individuals start to adapt to the generic HRTF [Wenzel et al., 1991].

Since the audio has to be rendered synchronously with the movements of the user, these have to be tracked. There are three common methods which can be used for tracking [Heller et al., 2014]: head tracking, device tracking, and body tracking. The movements of users are most accurately tracked by a head tracker and since headphones are used anyway, the required additional hardware can be mounted on the headphones or special headphones with integrated hardware can be used.

Head tracking is the most accurate tracking method.

There are three types of head movement angles: yaw, pitch and roll (Figure 1.1). In the common horizontal plane only rotational head movements along the vertical axis are needed (turning the head to the left and to the right), so only yaw movements are considered, for which a digital compass is sufficient. Localization in the median plane however involves pitch and roll, most importantly pitch, which is the head movement along the aural axis (tipping the head up and down). The roll movement is the movement in which the head is cocked to either a side and changes the relative elevation of the perceived sound sources (Figure 3.1). Rendering the elevation and

Three head movement angles: Yaw, Pitch, Roll.

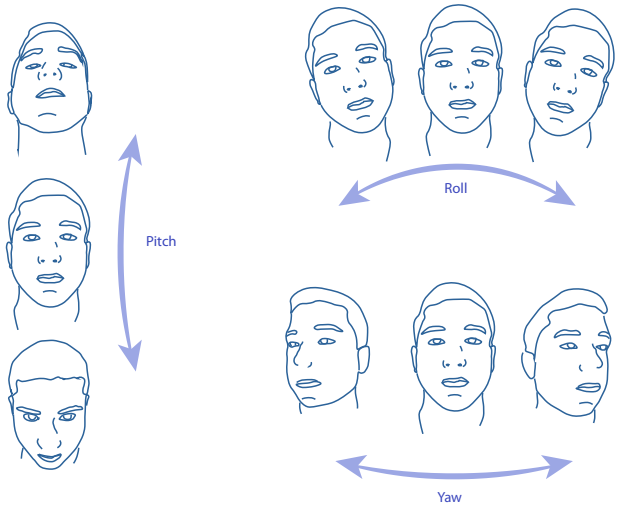

Figure 1.1: The three head movement angles (Euler angles): Yaw, Pitch and Roll

thus tracking pitch and roll movements might result in an increase of realism since the sound sources are not only perceived at ear level.
Localization performance of real and virtual audio sources in virtual audio spaces have already been investigated in several studies in the horizontal plane, since this is where most of the natural human orientation takes place. In this thesis we will investigate if the integration of the median plane into the virtual environment can improve the distinction of close positioned sources in the horizontal plane and if this could lead to an overall better localization performance and experience.

To examine this, we conducted two studies:

1. Determine the minimal audible angle (MAA) in the horizontal and vertical plane

2. Investigate if additionally simulating the elevation facilitates the localization of close-by sources in the horizontal plane and enhances the experience in the virtual audio environment

We will first of all discuss related work done in the domain of sound localization in real and virtual environments and their results, then we will describe the setup, implementation and procedure of our own experiments. At the end we will discuss the results, give a summary and look at possible future studies.

Chapter 2

Related work

Audio augmented reality systems have been studied for many years but have only recently gained importance with the emergence of mobile devices in which the hardware has shrunk drastically in size with an enormous increase in processing power. Early research have shown that navigation in a virtual environment can be successful even with minimal hardware. [Holland et al., 2002] for example showed with one of the first mobile audio augmented reality systems that this was possible by only using amplitude-panning and GPS measurements for the orientation and heading. With the increased computing power in mobile devices and development of modern sensors e.g. InvenSense MPU-9150[1] (used in the Intelligent Headset and KLANG:vektor, cf. chapter 3 "Setup"), more complex algorithms and higher resolution in perception are possible. Corona [Heller, 2014], for example, transforms the coronation hall in Aachen into a virtual medieval coronation feast by using virtual sounds emitted from different locations across the room. KLANG:technologies [2] provide a 3D in-ear monitoring system for musicians using audio augmented reality, so that every instrument is perceived at the actual location. The system also renders the elevation, since this is possible with modern technology, we want to investigate if this improves the virtual audio space.

Modern technology enables new possibilities.

[1] http://www.invensense.com/products/motion-tracking/6-axis/mpu-9150/

[2] http://www.klang.com

In this section we discuss some of the related work done in the field of audio augmented reality and sound source localization in real and virtual environments.

Subjects had to localize sources in real and virtual environments.

Localization of sound sources in virtual environments is less accurate than in real environments. The paper from [Bronkhorst, 1995] studies the several differences between real and virtual sound sources. For this, two tasks were conducted. In the first one the test subjects had to locate a continuous sound source and press a submit button when they located it, they were blindfolded and seated while the sources had to be located with head movements. There was a total of ten conditions of which two were held with real sources and eight were carried out with virtual sources. In the conditions for virtual sources half were done with individualized HRTFs which were measured in a previous task for each subject and the other half was done with non-individualized HRTFs. The sources were placed in 30 different positions varying in azimuth and elevation. In the second, the confusion task, the subjects had to indicate the region of the sources by pressing one of eight buttons of which each corresponded to one of the sources. The subjects also had to indicate whether the source was above or below the horizontal plane, which could be indicated through different shapes of buttons for the "above" and "below" positions. In this task the source emitted only a short sound.

Localization using individual HRTFs is as precise as for real sources.

The results from both tasks which were three experiments with 8 subjects conclude that localization of virtual sources with the use of individualized HRTFs is almost as accurate as with real sources with the condition that the sound is played long enough and head movements are allowed. With non-individualized HRTFs however localizing virtual sources is much poorer. Using a short broadband sound and disallowing head movements results in a much poorer localization for virtual sources compared to real sources, the confusion (front-back, left-right and diagonal) rate was much higher for virtual sources.

FRONT-BACK CONFUSION:
Front-back confusion are errors in perception that occur when a sound source which is in fact placed behind the user is perceived as it were in front of the user and the other way around. The same goes for left-right and up-down confusions.

Definition: *Front-Back Confusion*

In the vertical plane, localization was poorer for virtual sources. Furthermore localization of virtual sources was more accurate if the cutoff frequency was set to a value above 7 kHz. From this the authors conclude that, when using headphones, high frequency spectral cues are not simulated correctly, these cues are however important since they decrease the confusion rates and allow the subjects to locate the sound sources with fewer head movements and a higher accuracy. So this study additionally shows that the frequencies of the emitted sound have an impact on confusion rates. The authors finally also discovered that the subjects mostly begin with a horizontal (left/right) head movement which is then later followed by a vertical (up/down) head movement.

Frequency of the sound affects confusion errors.

[Wenzel et al., 1991] analyzed localization in a free field and a simulated free field over headphone condition with a non-individualized (generic) HRTF function. The purpose of this study was to determine if virtual acoustic displays are useful if only a non-individualized HRTF is used. They observed that experienced subjects had clearly lower confusion rates in both conditions than the inexperienced subjects. This concludes that there is a learning-effect when using generic HRTFs and that the higher confusion errors in virtual listening environments could be due to the unfamiliar listening conditions since confusion rates were also present in free field. For instance in [Wenzel et al., 1988] the authors observed that by using individualized HRTFs, if there were any performance issues, then it occurred in elevation while performance in the horizontal plane remained stable. The authors explain the reason behind this with certain missing or present acoustical features in the HRTF which are necessary for localization in the vertical plane. The paper finally suggests using additional cues for example visual cues to decrease

Learning effect for non-individualized HRTFs.

these errors. The influence of visual stimuli on localization performance is described in the following paper.

Perceptual fusion in the horizontal plane.

Localization performance on sound is not only influenced by audible factors, visual stimuli also have an impact on localization. A good example for this is the ventriloquism-effect. [Bertelson and Radeau, 1981] studied the perceptual fusion in the horizontal plane. Subjects had to locate sources by pointing with the hand, they used loudspeakers as audio stimuli and lamps as visual stimuli. There was one visual (subjects had to point at the light) and one auditory session (point at the sound direction) with a control trial with only auditory signals and a conflict trial with visual and auditory signals for each session. Subjects had to point at a target signal while ignoring a competing signal. Competing signals varied in distance from the target signal by 7, 15 and 25 degrees for each, once on the right and once on the left so that there were 6 trials. Results show approximate deviations of 4° for 7°, 6.3° for 15° and 8.2° for 25° separation.

Deviation of localization between audio and visual stimuli in the vertical plane.

The paper [Werner et al., 2012] investigates the visual-audio deviation in the vertical plane with binaural audio, for this the authors conducted two experiments. Virtual sound sources were placed in selected azimuth and elevation angles, while the visual stimuli which were LEDs placed in front of the user, ranging from top to bottom in a circle segment. The first experiment served to determine the participants experience with perceptual fusion and was divided into two sessions. In the first session in which the influence of visual stimuli on localization was investigated, participants had to answer if the audio stimulus was above, equal, or below the visual stimulus. The second session without any visual stimuli served to verify the perceptiveness. The results show that larger differences of audio and visual stimuli is more tolerable for the upper positions and upper lateral positions. The second experiment was used to verify and refine the results of the first experiment by using a different method. This time a laser pointer was used to indicate the location of the sources which were arranged in a tangent plane as were the visual stimuli. Visual and audio stimuli were combined in different orders. The results show that with

the sound source at 0 degrees horizontal and 0 degrees vertical there is a larger deviation for visual stimuli which differed in over 5 degrees. For the positions Horizontal: 20 Vertical: 0 (H20V0) as for H20V20 and H0V20 there are higher deviations for visual stimuli less than 15 and higher than 22 degrees. The two former papers show how additional stimuli can impact the perception of sound sources.

One of the minimal audible angle (MAA) experiments which involved decreasing the distance in the horizontal plane while increasing the distance in the vertical plane with real audio sources was conducted by [Perrott and Saberi, 1990]. In this study a number of 30 loudspeakers were fixed on a rotatable boom which could be rotated by 90 degrees. A reference source was randomly chosen from one of the 10 speakers in the middle of the array in each trial. Another source was then selected from the array and the subjects had to answer in a two-alternative forces-choice, if the second source was to the left or to the right of the reference speaker and with the array rotated at 90 degrees, if the second speaker was above or below the reference speaker. The reference source was in fact jittered between the trials so that any information which could have been extracted from the subjects was eliminated. If the response was incorrect the distance between the speakers was increased, if 3 successive responses were correct, the distance between the speakers was decreased. The boom was rotated at several angles, the MAA was calculated through the average at each angle position. The results of the study are very similar to earlier records with MAA values at 0.97 and 3.65 degrees for the horizontal and vertical plane respectively. With the boom rotated at 0 to 60 degrees the horizontal MAA values are constant but at above 60 degrees, rotating the array had no effect on localization resolution. The authors state that post examinations indicate that the study only works with binaural processing since monaurally, the MAA values for the vertical plane increased drastically.

Determining the MAA by decreasing the horizontal distance and increasing the vertical distance.

The paper [K. Saberi] analyzes the MAA values with a very similar setup as the previous research but on the lateral and dorsal plane of the listener. The resulting graph is similar to the previous study for the dorsal plane, in the

MAA in lateral and dorsal plane.

lateral plane however, the function is the inverse of the dorsal plane. The authors discuss that head movements are essential for an improvement on localization, especially in the lateral plane.

Determining the MAA in a virtual environment.

In the paper from [Wersényi, 2007], listening tests are performed to determine the minimal audible angle in an HRTF-based environment, i.e., with virtual audio. The authors use a generic HRTF which originates from a good listener. They used untrained subjects for their tests and state that this is common in listening tests since their localization skills are poorer than trained subjects, as stated earlier, there is a learning-effect. A loudness listening test was done before the main test and headphone errors were considered as well. The former properties are necessary since localization depends on loudness (externalization effect) and a-priori knowledge (learning-effect). Localization also depends on duration and signal frequency that is why this study uses 3 different broadband signals. The authors state that signal bursts with a duration over 250 ms are optimal for localization. Furthermore the optimal frequency range lies below 1000 Hz and above 4000 Hz. A listening-test was performed for the horizontal and vertical plane independently. A reference source was chosen at a certain angle. Subjects had to tell if they could distinguish between the position of the reference source and a candidate source which was moving back and forth of the reference source. The subjects had to answer in a 3-category-forced choice, i.e., answers yes, no or not sure. The position of the candidate source at which the subjects could differentiate between the two sources was selected as next reference position. The obtained MAA values are generated through the average of all subjects, and differ for each signal type with the best results achieved with white noise.
Averages are 7-11° and 15-24° for the horizontal and median plane respectively. This data is comparable with other studies, there are some larger differences between the best and worst localizer. Localization of virtual sources in the median plane as for real sources is much weaker than in the horizontal plane. The authors suggest not to use sources in the vertical plane because some of the subjects could not distinguish any of the sources separated

by elevation.

In [Jan Abildgaard Pedersen, 2005] the minimal audible angle was compared between real and virtual sources in part of a 3D auditory display system used for fighter aircrafts in which alarm signals (virtual audio sources) should sound from a certain direction of a real object. The authors determined the azimuth and elevation errors independently, they used 58 source positions which were either on the right side or on the left side of the participants, instead of 107 sources in both sides to reduce time. The sound sources covered the whole radius around the subject. White noise was used as sound sample with different durations of bursts. Half of the subjects were fighter pilots, the other half were civil persons. 16 of the 58 sources were selected as real sources, so that there were several sessions with different bursts and source types (real or virtual). The order of the sessions was randomized for each participant. A curtain was placed around the subjects to hide the speakers of the real sources. For the virtual sources a head-tracker was used. The procedure was simple, the subjects heard the sound from a certain direction and had to point with a toy-gun which was equipped with a tracker at the direction of the source. The results show the error for the real sources was at ca. 10° for azimuth and 12° for elevation while for virtual sources the values were 14° and 24° respectively. This paper clearly shows the difference of localization performance between real and virtual sources, especially in the vertical plane and also shows an interesting use case for audio augmented reality systems.

Comparison of the MAA between a real and virtual environment for a 3D Auditory Display.

[Mariette, 2010] analyzed how the head-turn latency and rendering method in a virtual audio environment effect navigation performance. The main experiment was carried out on the outside in an open field, subjects had to press a button to start the experiment, a sound that was binaurally rendered was played at a certain position and the subjects had to walk towards the sound until it stopped playing. The subjects had to rate the perceived sound source by stability. The participants were encouraged to use head-turns to locate the sources and to look upright while walking towards the source.

Effects of head-turn latency and rendering method on navigation performance.

In a previous pilot study the optimal capture radius around the sources was investigated and different head-turn latencies were used. The author analyzed the distance efficiency with the results showing that a higher capture radius results in a higher distance efficiency with the conclusion that a capture radius of 2 meters was sufficient. Thus the author used this value in the main experiment. The results on head-turn latency indicate that higher head-turn latency had no effect on distance efficiency but lead to decreased stability (depending on the rendering method) while independently of the rendering method, head-turn latencies lower than 200 ms with a total system latency of 376 ms did not affect performance significantly. Anyhow navigation of the participants was successful even with high localization errors and performance issues.

Chapter 3

Setup

Most of the audio augmented reality systems only simulate the horizontal plane of virtual sound sources in a virtual environment, so that all the sources are perceived at ear height. We want to figure out if additionally rendering the audio in relation to pitching and rolling the head helps discerning two sound sources placed close to each other and if this can lead to a better experience. To simulate the elevation, additional sensors in the tracking system are needed, therefore head-tracking is the only option since pitching the head changes the perceived height of sound sources and tilting the head changes the relative position of the sources in the vertical plane (Figure 3.1). If we can improve the localization of virtual sound sources by simulating the elevation, then one could place the sound sources closer together thus enhance the experience, leading to a more realistic virtual audio environment.

Does the simulation of the median plane have an impact on localization performance and presence in the virtual environment?

To determine this we conducted two studies. The first study served to figure out the minimal audible angle in the horizontal and vertical plane, i.e., virtual sound sources were placed at different directions and different heights in two separate tasks respectively. The purpose of this study was to see how well users could discern two sound sources separated in the same plane in different angular regions of the respective plane, this should give us information for the initial source placement of the second study.
In the second study we wanted to figure out if simulating

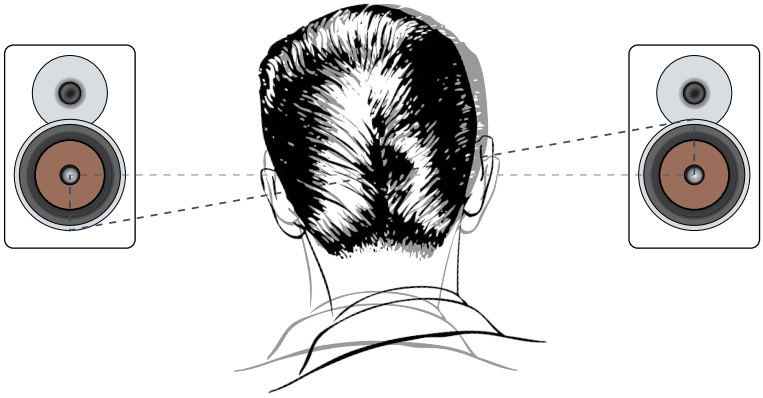

Figure 3.1: Rolling the head to the left or to the right changes the relative elevation of the perceived sound sources, thus simulating elevation might help discerning two sources even if they are placed at the same height

the elevation, i.e., rendering the audio additionally in relation to pitch and roll head movements, reduces the minimal audible angle in the horizontal plane. For this, a localization test was carried out. Both studies were lab-based.

3.1 The Intelligent Headset

The most accurate way to track the movements of a user is by using a head tracker [Heller et al., 2014] and since we have to track head movements, this is the only option. For this to function, additional hardware is needed, this hardware can be integrated or mounted onto the headphones. For our experiments we used the Intelligent Headset (IHS) from Jabra[1]. It is a headset with various integrated sensors: GPS, compass, gyro, and accelerometer which allow tracking the movements of a user. Furthermore it can be used wired or wirelessly via Bluetooth which can be very useful in navigation tests since wires might influence or disturb the orientation performance, this however comes with the cost of latency. Wired, the IHS has a specified latency of

The IHS has the required sensors to track all the head-movements.

[1] https://intelligentheadset.com/

100 ms which is clearly below the 372 ms stated in [Mariette, 2010]. Head Orientation changes were transmitted via Bluetooth at around 40 Hz while audio was transmitted through the wire.

3.2 The Klang App

Rendering of the audio was done by an App called KLANG:app (Figure 3.2) from KLANG:technologies which runs on iOS devices. Klang is a company which provides a 3D in-ear monitoring system called KLANG:fabrik for musicians which is why it has a very low latency of approximately 10ms. Through binaural rendering, their technology provides a natural, high quality and transparent 3D sound for live performances on the stage. It uses a generalized HRTF for rendering.

Rendering with very low latency at about 10ms.

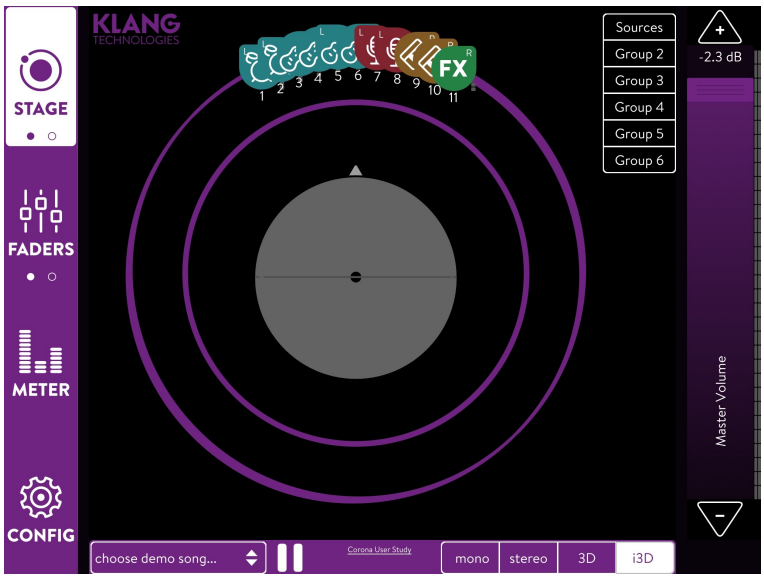

Figure 3.2: KLANG:app resp. KLANG:kern Graphical Interface: Here the 11 sources used in the second experiment are shown which are placed with a 5° spacing in azimuth, ranging from -25 to 25 degrees.

In the second experiment another version of the KLANG:app called KLANG:kern with the same rendering engine as the KLANG:app was used for the audio rendering. It was released later in this occasion and was better suited to our requirements because other than the KLANG:app, KLANG:kern allowed us to load several audio files as sources at different positions at once in the virtual space using an XML file while the KLANG:app allows loading just one file. Note that if we use the term Klang App, we refer to both versions.

"Imagine you are singing – live on stage. Now you turn around to face your band. Wouldn't you wish to hear every single instrument from exactly where you see it and not just from your In-Ear? An In-Ear sound that is as natural as hearing without In-Ears?" (klang.com)

The App can be controlled remotely by OSC messages.

The rendering capabilities of KLANG:fabrik are integrated into the Klang App which allows a flexible placement of virtual instruments (sound sources) across the room, creating an individual mix. The KLANG:fabrik as well as the Apps can be controlled remotely by using the Open Sound Control[2] (OSC) protocol which permits most importantly (for our experiments), orientation updates and source positioning in the horizontal and vertical plane. The OSC protocol is a protocol which allows sending network messages, it was originally designed for computers, sound synthesizers and various other multimedia devices to communicate with each other through modern networks using UDP/IP.

Definition: *Quaternion*

QUATERNION:
Quaternions are a set of four components equal to a four-dimensional vector-space. This four-dimensional number system extends the complex numbers and is used to describe three-dimensional space.

By using the Cocoa OSC library (by Daniel Dickinson) the test-apps which we created for both tests were able to send the quaternion values to the Klang App via OSC messages. These messages are sent through the network over wireless

[2] opensoundcontrol.com/

LAN. OSC messages use an URL-style symbolic naming scheme. For example the URL-scheme to set the position of a sound source remotely in the Klang App looks like the following:

```
/KLANGfabrik/user/ChannelPosition , 'ifffs'
, <ChannelNumber>, <X>, <Y>, <Z>, <UID>
```

The message starts with an URL address pattern, followed by the channel number, an Integer (i) of which the position is going to be set, this is then followed by the cartesian coordinates X, Y, and Z of the position which are float (f) values and at last UID which is a String (s) with the User ID. This set of data can then be sent as an OSC packet. Since the Klang App currently has a remote command which only allows changing the cartesian positioning of the sound sources, the elevation and azimuth angles had to be converted to the equivalent Cartesian values.

3.3 Implementation

The test-apps for our experiments were all mac applications and programmed in Objective-C. As previously mentioned, we used the Cocoa OSC library to send OSC messages. The test-apps served as Graphical User Interfaces and controlled the whole test procedures.

Implementation of test-apps and a converter App.

Since the Intelligent Headset (IHS) cannot directly send data to the KLANG:kern and as the latter is capable of receiving OSC messages, we created an iOS App (IHS converter). The IHS converter which runs in the background while using the KLANG:kern is capable of receiving and processing the raw data of the IHS sensors and convert them to quaternions which can be then processed by the KLANG:kern. Furthermore additional features such as calibrating the compass and enabling resp. disabling the simulation of elevation were integrated into the IHS converter App so that for the second experiment, the test-app was able to send OSC messages to the IHS controller App which would execute the corresponding function. Additionally, the head orientation data could be sent to a second IP address for logging.

3.4 Technical Setup: First Experiment

The goal of the first study was to determine the minimal audible angle in the horizontal and vertical plane using the rendering capability of the KLANG:app running on an Apple iPhone 5 with iOS 8. For the test itself a mac test-app with a simple user interface was created in which the participants could start the test and select the answers. The test-app was designed to send the speaker positioning to the KLANG:app of a reference source via OSC and switch to a second comparing source. The participant should then select an answer and the app reacted accordingly.

Non-continuous sound used for the listening test.

We used the Intelligent Headset as headphones, orientation tracking however was not necessary in this experiment. The sound file in the KLANG:app could easily be replaced, so we used the same non-speech beacon-drum sound as in [Heller et al., 2014] as sound sample. It is a more natural sound than white noise which is often used in hearing tests because it covers up the complete hearing range, however white noise is a sound that does not occur in natural listening environments.

3.5 Technical Setup: Second Experiment

The goal of the second study was to determine if the angle separating two virtual sound sources in the horizontal plane is lower in an environment in which the elevation is simulated compared to an environment in which only the horizontal plane is simulated and if this causes a better experience.

IHS converter App to connect the KLANG:kern with the IHS.

In the setup of the second experiment we used the Intelligent Headset as headphones and as primary motion tracker. The KLANG:kern was used for the audio rendering. The App ran on an iPad Air 2 with iOS 8. The IHS converter App ran in the background allowing us to use the Intelligent Headset and enable the formerly stated additional features (3.3 “Implementation”).

We used the KLANG:vektor which is the original motion tracker used in the 3D in-ear monitoring system by Klang

as a secondary motion tracker. The KLANG:vektor would have allowed a direct connection to the KLANG:kern instead of using a converter app for the motion tracker of the Intelligent Headset, thus reduce the latency since the system is designed for live 3D in-ear monitoring on the stage. However due to technical complications it was not capable of sending its data to two IP addresses, i.e., to the logger and KLANG:kern simultaneously, so that the IHS which has the same motion sensor (InvenSense MPU-9150) the KLANG:vektor uses, was used as primary tracker instead. The motion sensors of the KLANG:vektor were mounted onto the Intelligent Headset so that we could log the data in addition to the IHS data and compare the results later on. The KLANG:vektor was connected to the iPad with a standard 3.5mm stereo audio cable and to the network via its integrated Wi-Fi sending its data to only the logging device.

IHS and KLANG:vektor use the same sensor.

A second test-app had to be developed which had similar functionality as for the first experiment. This test-app was also capable of sending OSC commands based on the executed actions. The test-app with its graphical user interface allowed us to select the answers of the participants and was responsible for the whole test procedure by randomly selecting the sources which had to be played at different positions using latin squares. Selecting the active sources was done in the same way as in the previous test with an OSC command that was sent to mute resp. unmute the source. Furthermore all the important information could be monitored with the test-app, it logged all of the answers and orientation measurement which were transmitted from the IHS converter App to the test-app via OSC messages. The IHS converter simultaneously sent the tracking data to KLANG:kern for rendering and to the test-app for monitoring through two different IP addresses. The quaternions received from the motion tracker were then converted to euler angles by the test-app and logged for later analysis.

Test-app controlled the procedure.

For the wireless network, a stage router was used, the test-app, KLANG:vektor and iPad with the KLANG:kern were all connected to the router so that OSC messages could be transmitted to each other.

Card-board tubes as visual sources.

As visual sources, we used 17 numbered cardboard tubes with two different heights (140 cm and 70 cm depending on the condition cf. Section 4.2 "Second Experiment: Localization Performance"). The tubes were placed at a distance of 2 meters to the listener ranging from -40° to 40° (Figure 3.3). In the listening test we tested the angular

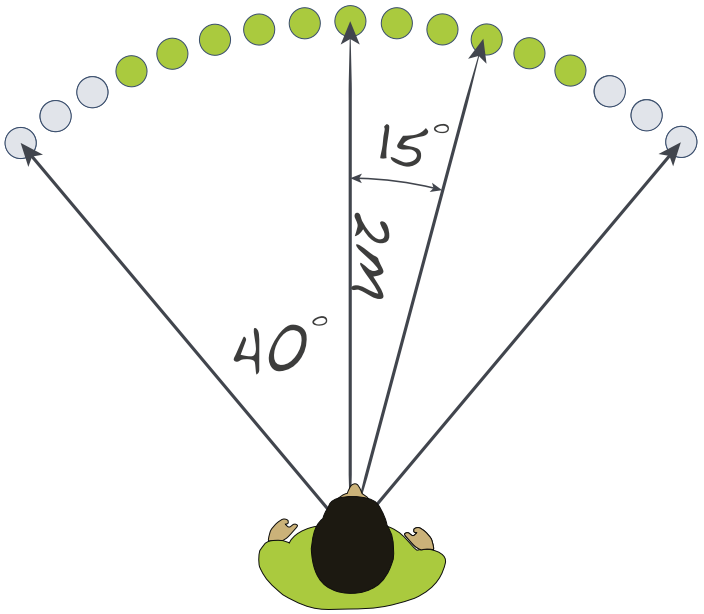

Figure 3.3: Placement of the 17 cardboard tubes in the setup of the localization test. The six sound sources on the outer positions marked in grey, were used as placeholders to avoid artificially limiting the range, the actual sources (in green) were marked as active by a loudspeaker for the corresponding angular spacing of 5°, 10°, 15° and 20°.

spacing of 5°, 10°, 15° and 20°, we marked the respective sources as active by placing speakers on top of the tubes (Figure 3.4), i.e., every second, third or fourth tube from the center was marked as active respectively. We chose a minimal spacing of 5° since the MAA in the first study at 0° (at the center) was about 4.8° (Chapter 4 "Evaluation"). The actual sounds were played in the range of -25° to 25° (Figure 3.2), the additional tubes were placed to avoid limiting the range towards the outer sources. Since the sources were all clearly in front of the listener, front-back

confusion errors were non-existent.

Figure 3.4: Sources (cardboard tubes) were numbered and active sources were marked by placing a small loudspeaker on top of the cardboard tube

We used a continuous monologue of a male voice so that we could replicate a natural use case e.g. the situation at a museum. In addition the sound is continuous, so that the duration of the signal is optimal since [Bronkhorst, 1995] and [Wersényi, 2007] suggested that short sounds result in a poorer localization. Furthermore the frequency of the male voice is below 1000Hz, thus lies in the optimal frequency range stated in [Wersényi, 2007].

Continuous monologue voice used for the localization test.

Chapter 4

Evaluation

4.1 First Experiment: Minimal Audible Angle

In the first study we wanted to determine the minimal audible angle (MAA) which is the smallest angle separating 2 sound sources so that they can be distinguished. This is important for the second experiment since we need to know the characteristics of the the Klang App's rendering and HRTF. The MAA listening test consisted of two tasks, a test in the vertical plane and a separate test in the horizontal plane.

Determine the horizontal and vertical MAA.

4.1.1 Participants

5 male users between the age of 22 and 25 were randomly selected as test participants for the listening test with the procedure described in the following section. The participants were verbally informed about the procedure and the KLANG:app was shortly demonstrated to them so that they could get a better understanding for the upcoming test. The azimuth test took about 10 min and the elevation test was about 5 min long while in overall it took about 20 min (including a short break between the tasks). None of the participants had any experience with virtual audio and

neither of them had ever taken part in a listening test. Inexperienced participants were chosen intentionally since their localization skills are usually inferior to experienced listeners which is justified through the ”learning effect” (Chapter 1 “Introduction”, Chapter 2 “Related work”).

4.1.2 Procedure

The listening test was performed using the Mac test-app mentioned in the previous section as a user interface and the KLANG:App as the renderer (3.4 “Technical Setup: First Experiment”). The procedure was based on the Up-and-Down transformed response rule (UDTR) in [Wetherill and Levitt, 1965] but had to be adapted to shorten the test to avoid aural fatigue of the participants. The participants were seated in a quiet room in front of the computer and the volume was set to a comfortable level. As soon as the participants started the test by pressing the start button in the user interface of the elevation task, a reference source was randomly selected from a set of angles. The set of angles for the test in the vertical plane contained the following values: -40, -20, 0, 20, and 40 degrees with 0° being at ear-height. These numbers were chosen based on previous research results of various studies. With the above values we made sure that we could cover the range as good as possible without increasing the duration of the test to a fatiguing extent.

Since the KLANG:app cannot receive any stop or pause commands, the sound had to be muted before the test and was unmuted as soon as the test was started. The reference sound was played for about 5 seconds and was immediately muted for 1 second. Then the source was randomly moved above or below the reference source with an initial distance of 4° to the comparing source. This value was also chosen based on the results of previous MAA studies with real sources in which the MAA was about 4° in the vertical MAA. The sound was then played in the second position, for also 5 seconds. The participants had to answer in a two way forced alternative choice if they could hear a difference between the 2 source positions, so the only possible answers were Yes and No. If the participant gave a negative

Sources were muted resp. unmuted and moved further apart after negative answers.

answer, the distance between the 2 sources was increased by 4°. The run was then repeated with the reference source at same position as before while the comparing source was positioned at 4° further apart. If the participant gave a positive response, the distance was decreased by 2°, nearing the MAA at this position gradually. Giving a second consecutive positive answer confirmed the current distance as the MAA and resulted in randomly choosing a new reference angle from the set and repeating the procedure with the initial distance starting at 4°. A negative answer after a single positive response was treated the same way as a single negative answer. The test had a minimum and maximum value of -70 and 70 degrees respectively, if these were exceeded a new reference source was selected and the run was marked as failed, if the participant had given a positive answer before failing, this value was taken as the MAA at this position. The test was completed when all of the reference angles of the set were used. A short break was taken between the two tasks. The order of the tasks was alternated between the participants.

Two tasks, one for elevation and one for azimuth respectively.

For the other task, the azimuth MAA test, the procedure was the same but with a different set of values: -60, -40, -20, 0, 20, 40, 60 degrees with 0° being the position right in front of the participant. The distance between the two positions was increased by 2° for negative answers and decreased by 1° for positive answers. These values are as well based on the MAA results for real sources from previous research for the same reason as stated in the elevation test above. The answers of the participants, reference source position, comparing source position, distance threshold and a user ID were logged in a file so that the data could be processed later on.

4.1.3 Results

The results are similar to other studies, while there are high deviation especially in the wider angles. We calculated the mean values for each reference position. In the horizontal MAA test (Figure 4.1), the results are mixed for the different ranges while the center at 0° has clearly the lowest minimal audible angle at 4.8° and the lowest deviation be-

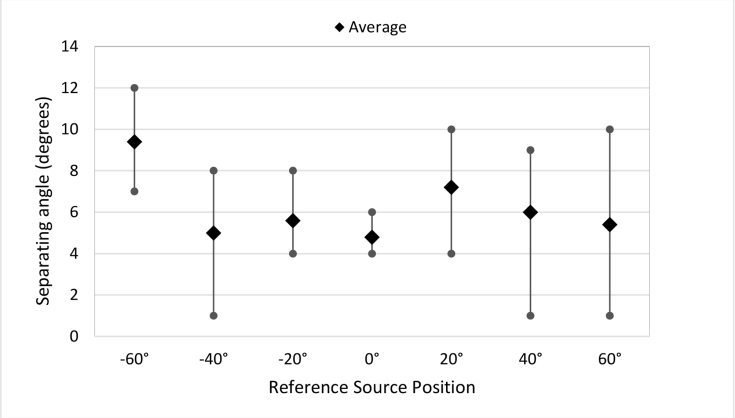

Figure 4.1: Representation of the horizontal minimal audible angle, the error bars correspond to the minimum and maximum values of the participants while the hashes represent the mean values

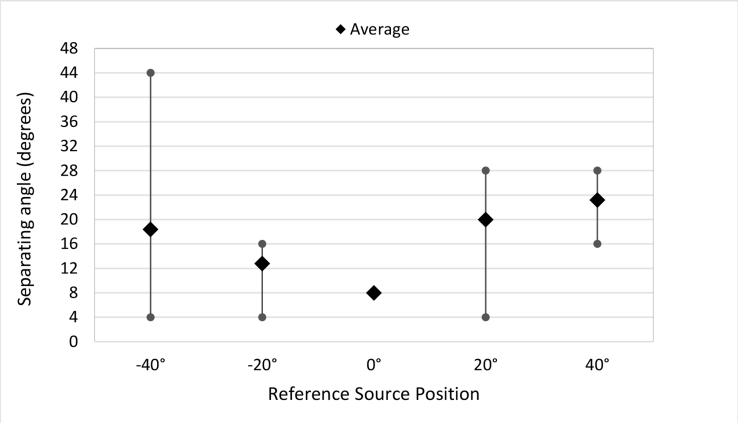

Figure 4.2: Representation of the vertical minimal audible angle, the error bars correspond to the minimum and maximum values of the participants while the hashes represent the average values

tween individuals. The mean values of the other areas except the furthest one on the left at 60°, do not differentiate that much, the minimum and maximum values of the participants however go further apart as the reference angle increases.

The test in the vertical plane (Figure 4.2) gives a clear pattern, at ear-height, at 0°, the minimal audible angle coincidentally for all of the participants was the same at 8°. The MAA and deviations increase with higher reference angles. The calculated mean for all reference positions thus the MAA is $M = 6.2°, SD = 2.77$ for the horizontal plane and $M = 16.48°, SD = 10.02$ for the vertical plane while the specified MAA of the HRTF used in the Klang rendering engine is 1° in the horizontal plane and 5° in the vertical plane.. All of the participants were at some point able to distinguish the two compared sources in both tasks.

4.1.4 Discussion

In comparison with the study of [Wersényi, 2007] in which the course of the tests was similar, the MAA values in the horizontal plane in our experiment are lower than the values in the paper, while in the vertical plane, they are similar. Furthermore one must consider that the authors in the paper used broadband noise which should provide better results than the drum sound we used, but still the values of our experiment are lower. Analysis of the drum sound in [Heller et al., 2014] shows that it also covers a broad frequency range. One must however expect differences in comparison with other tests in this field since the results depend on many factors such as sound samples, rendering mechanism, HRTF, methodology etc.

The KLANG:app has a high Fidelity audio rendering.

[Middlebrooks, 1999] for example got values of 17.1° with non-individualized HRTFs and 14.7° with individualized HRTFs in the horizontal plane while [R. L. McKinley, 1997] had very similar minimal audible angle values to our ones of 5° for 0° azimuth and constantly growing MAA values with an increasing target direction angle towards 90°. At 30° the MAA was about 5.5°, at 60° it was at about 8° and at 90° it reached ca. 15°.

In the vertical plane for example [F. L. Wightman, 1989] determined the vertical MAA values in the horizontal front, side and back position in low, middle (at ear height) and high elevations. Low, middle and high elevation were defined as -36 to -18 degrees, 0 to 18 degrees and 18 to 36 degrees respectively. In the front position for example, the

mean values were 20.4° for low elevations, 17.9° for middle elevations and 25.2° for high elevations.
We conclude that the KLANG:app/KLANG:kern has an exceptionally high standard rendering engine since the MAA values of our test and the specified MAA values by Klang are low compared to other studies and rendering engines.

4.2 Second Experiment: Localization Performance

Three conditions for the localization test.

In the second study we wanted to figure out if simulating the elevation can improve the localization resp. discernment of sources in the horizontal plane. For this we conducted a localization test in which the participants had to localize and name the active sources. This test consisted of three conditions. In the first two conditions the sources were all placed at the same height, in one condition the elevation was simulated, i.e., all three head movements (Figure 1.1) were simulated (from here on referred to as *elevation*). In the other condition only horizontal head movements were tracked, i.e., only yaw movements. (referred to as *flat*). Finally in the third condition the sound sources and the corresponding visual sources were placed at different heights and the simulation of elevation was enabled (referred to as *physical*).

4.2.1 Participants

22 users between the age of 22 and 40 with an average of 28 years participated in the study. Three of them were females and the other 19 were males. None of the participants reported to have a hearing disorder neither did any of them have problems with spatial hearing. About half of the participants had experience with audio augmented reality. The listening test took about 30 minutes in total with short breaks between the rounds and tasks.

4.2.2 Procedure

The test was conducted in a quiet room. Rendering was done using the KLANG:kern App. The whole procedure was controlled by a Mac App (Section 4.2 "Second Experiment: Localization Performance"). The test-participants had to stand upright on a marked spot on the floor, in front of the arc of cardboard tubes (Figure 4.3). They were wearing the Intelligent Headset. Since they were standing on a fixed position, the headphones could be wired without influencing the movements and thus reducing latency. The procedure was verbally explained to the participants. They were given a training session before the actual test in which they could localize some random sources. All actions were done verbally, i.e., the participants gave the order to start the test and were instructed to clearly speak out the number of the source they thought was playing.

Figure 4.3: User standing in front of the arc of cardboard tubes while performing the listening test, in this picture every third tube is marked as active, i.e., an angular separation of 15° was used. The short tubes (70 cm) used for the third condition were half the size of the tall tubes (140 cm).

There were three conditions in total from which the first two consisted of 4 rounds each, a round corresponded to

Including and excluding simulation of elevation in the first two conditions.

the angular separation which were 5°, 10°, 15° and 20°. The first two conditions and rounds were randomly chosen using latin squares. From the 11 actual sources, there were 10 repetitions per round. The third condition consisted of one round with an angular separation of 5° so that in total there were 90 trials per user. The repetitions, i.e., the selection of sound sources was also randomized using latin squares so that the whole procedure was counterbalanced. Before each trial the participants had to keep their head straight, i.e., look towards the tube in the center, so that the compass could be calibrated if necessary since the compass sensor of the IHS was drifting slightly. As soon as the participants gave clearance the test was started, a random source was selected and the sound was played. The participants had to locate the sound by only using head movements while standing upright. When the source was located, the participants responded with the number of the source and the source was muted, then the procedure was repeated. Before each round the active sources were marked by a loudspeaker. After all 4 rounds for the first task were completed, the participants were given a short break and had to fill out a questionnaire (Figure A.1), there was one questionnaire for each of the first two conditions. After the short break the test was resumed and the remaining rounds were completed.

Description of the procedure for the user.

In the third condition we placed the sound sources as well as the corresponding cardboard tubes at two different heights. From the center, every second tube was 140 cm high, i.e., at approximate ear height, which was the height used in the previous 2 conditions. Every other source was placed below at a height of 70 cm. In the KLANG:kern (Figure 4.5) the virtual sound sources were placed at 0° and -45° of elevation respectively (Figure 4.4) . We wanted to investigate if placing sources at different heights has an influence on the accuracy of localization. In this condition only an angular separation of 5° was used since we expected that for larger angular separations, the sources would be discernible anyway. Tracking the elevation was of course enabled. The participants were instructed to only give the number of the playing source since the vertical position of the sources was clear.

Sources placed at different heights in third condition.

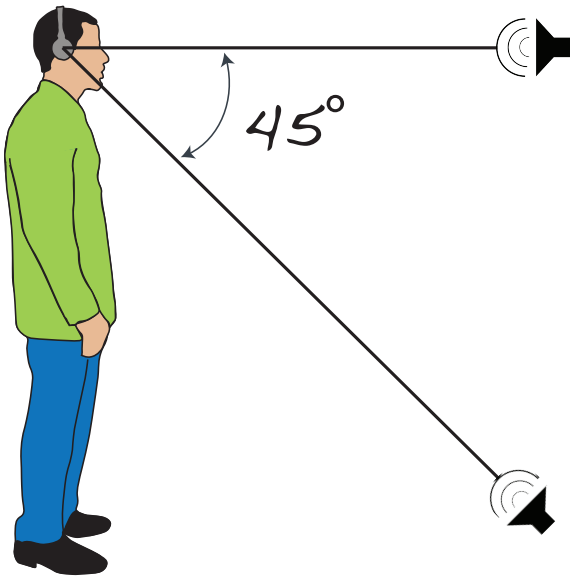

Figure 4.4: In the third condition of the localization test, every second sound source was placed at -45° while the other sources remained at 0° ear-height. This corresponds to the short (70 cm) visual source and tall (140 cm) source respectively

4.2.3 Results

Since we recorded the current source that was played and the answers of the participants, taking a look at the accuracy of the given answers shows us that even though some participants achieved a recognition percentage of up to 70% for the angular spacing of 5°, in general, localization accuracy for this degree of separation was poor in the *flat* and *elevation* condition (*elevation*: M=29%, SD=46%, *flat*: M=30%, SD=46%). In the third, *physical* condition in which the sources were placed at different heights, no significant difference is seen (*physical*: M=33%, SD=47%, F(2,657)=0.37, p=0.69). By only considering the two heights of the sources used in the third condition, the rate of telling if a source was up or down correctly was 56% in average. Increasing the degree of separation, also increases the rate of accuracy, however there is no significant difference between *flat* and *elevation* (cf. Table 4.1)

5 degrees angular spacing is too small for successful discrimination.

Figure 4.5: Sources in the KLANG:kern placed at different heights in the third condition, at 0° (ear-height) and below at -45°, corresponding to the 140 cm and 70 cm visual source (tube) respectively.

	Condition	Angle 5°	10°	15°	20°
Recognition rate	*elevation*	M=29% SD=46%	M=62% SD=49%	M=76% SD=43%	M=86% SD=34%
	flat	M=30% SD=46%	M=64% SD=48%	M=80% SD=40%	M=85% SD=36%
	physical	M=33% SD=47%			
Task compl. time	*elevation*	M=9.65s SD=5.31	M=9.18s SD=10.11	M=6.24s SD=3.57	M=5.43s SD=3.14
	flat	M=9.02s SD=3.69	M=7.21s SD=3.46	M=5.62s SD=3.91	M=5.35s SD=3.14
	physical	M=12.64s SD=8.50			

Table 4.1: Percentages of correctly identified sources and task completion time with standard deviation by angular distance for all 3 conditions

Furthermore when the participants had prior experience, the recognition rate, as seen through repeated measures ANOVA (F(1,1743)=8.02, $p < .005$), was higher than without, since there is a learning-effect. With an angular spacing of 10° for example, the recognition rate which is 56% (SD=49%) without experience climbs to 70% (SD=46%) with prior experience. The significance of this difference can be proven by a post-hoc t-test ($p < .005$).

Learning-effect.

Regarding the task completion time (cf. Table 4.1) , participants took much longer to localize the sound sources in the *physical* condition (M=12.64s, SD=8.5) in which only the 5° spacing was tested, than in the other two conditions at the same angular distance in which the sources were all placed at the same height (*elevation*: M=9.65s, SD=5.31, *elevation*: M=9.02s, SD=3.69) (Figure 4.6). The learning-effect can be observed again in the *physical* condition, participants could locate the sources quicker if they had prior experience (M=7.05s, SD=3.9) and were slower without experience (M=8.2s, SD 8.2).

Slower localization in third condition.

We calculated the root mean square (RMS) for the three head-movement angles Yaw, Pitch, and Roll to investigate how much the participants moved their head in each of the three angles (Figure 4.7). We compared the RMS angles received for the three conditions with 5° spacing. The results show that the yaw head-movement of the participants is very similar in the *physical* (M=18.96°, SD=5.04) and *elevation* (M=18.54°, SD=2.64) condition while in the *flat* condition they turned their head slightly more (M=20.26°, SD=6.38). We performed repeated measures ANOVA with user as random factor, this shows that the condition had a significant effect on Roll (F(2,42)=6.4287, $p = .0037$) and Pitch (F(2,42)=4.2739, $p < 0.05$). Post-hoc t-tests with Bonferroni correction show that in the *physical* condition, participants rolled their head more than in the other two conditions ($p < .01$). The RMS values of the pitch movements only differ significantly between the *physical* and *flat* condition, this shows that participants subconsciously pitched their head while localizing sources when all three angles were simulated although they did not notice any difference. For the other separation angles, there was no significant difference in RMS angles. The questions in the ques-

Participants nodded their head more in the third condition.

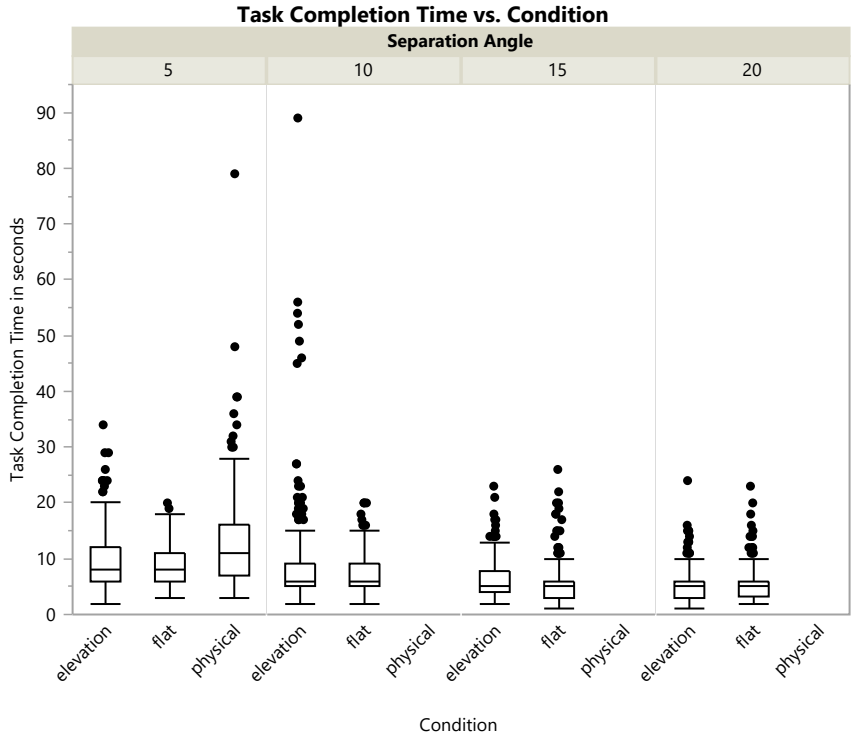

Figure 4.6: Task completion time for each of the four angular spacings vs condition, note that in the third condition (*physical*) only 5° angular spacing was tested

No significant difference in perception between conditions.

tionnaire (Figure A.1) were answered on a 5 point Likert scale with 1 being the best and 5 being the worst. The median ratings show only minimal differences. The question *How much did your experience in the virtual environment seem consistent with your real-world experiences?* received a slightly better rating for the *elevation* condition than for the *flat* condition (Mdn=2, IQR=2 vs. *flat*: Mdn=3, IQR=2.5). When asked how natural the interactions with the environment were, the participants rated both conditions equally (Mdn=2, IQR=2). The responsiveness of the virtual environment was rated very positively (Mdn=1, IQR=1) and the participants seem to have quickly adapted to the system for both conditions (*elevation*: Mdn=1, IQR=1.5; *flat*: Mdn=1, IQR=2).

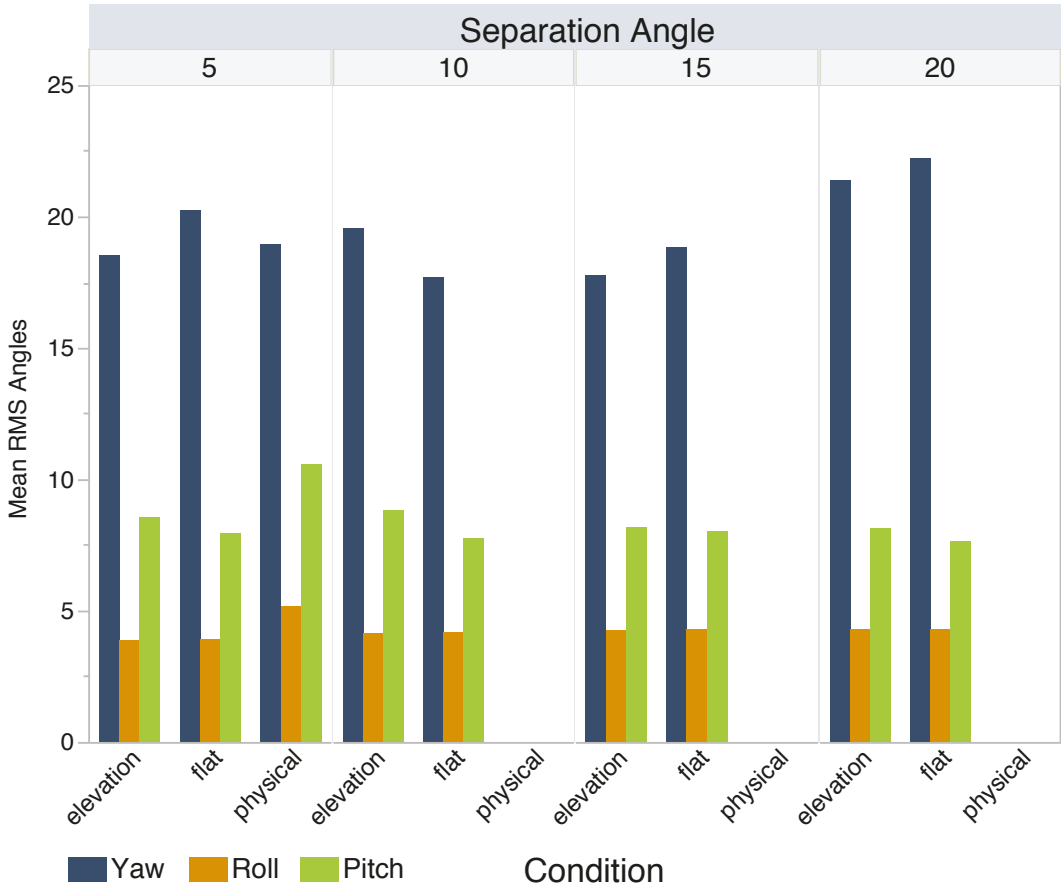

Figure 4.7: The average RMS angles for yaw, pitch, and roll by condition and source separation angle. While the RMS pitch angles are slightly higher in the physical condition, the difference to the elevation condition is not significant

4.2.4 Discussion

Compared to other systems [Heller and Borchers, 2015], the angular distance in our experiment is about 50% smaller. Surprisingly, there was no significant difference between only rendering yaw and rendering all three head angles. There was no notable improvement in localization performance as we had expected. The users reported that there was no noticeable difference between the *elevation* and *flat* condition. During the second run, most of them were confused and asked if there was any difference. Just 2 of the 22 participants noticed a minimal difference, one claiming that *flat* was easier, while another one thought

No improvement in recognition rate with elevation.

that *elevation* sounded more natural. Note that participants did not know about the difference in the first two conditions. Even though users did not notice any difference, they seem to have moved their head more in the *elevation* condition which shows that simulating the elevation does have a slight impact on the way users localize virtual sound sources after all. Some users reported that with their eyes closed, the position of the source that was playing was clear and that when they had to chose between visual sources, there was a certain deviation from their expectancy. This could possibly be a consequence of the ventriloquism effect [Bertelson and Radeau, 1981, Werner et al., 2012]. Some users also reported that the sound was actually coming from in-between the marked active sources, this could be due to the fact that the compass of the IHS was drifting moderately. Furthermore some users stated that they were missing room acoustics, i.e., distance of the sources to the listener and navigating towards them (cf. 5.2 "Future work").

Most users did not hear any difference in height.

Placing the sources at different heights in the third condition astonishingly also did not improve the discernment of sound sources. In the third condition, high and low placed sources were vertically separated by 45° and the participants could tell if the source was up or down correctly by only 56%, this shows that there was either an up-down confusion or participants just did not hear any difference in height. Most users reported that they did not hear a significant difference in height, even though the MAA discovered in the first study was 16,48° in average with the worst result at 44°. Some of the users even nodded their head extremely to hear a difference.

Chapter 5

Summary and future work

In this thesis we investigated if the simulation of elevation in mobile audio augmented reality systems has an impact on localization performance and the overall perception of the virtual environment, precisely we explored if there is an improvement in the discernment of virtual sound sources placed close to each other when all three head angles are simulated and when sources are placed at different heights. We conducted two studies. The first study was a simple listening test in which the participants had to tell if they could hear a difference between two sources which varied in angular distance to each other in the horizontal and vertical plane respectively. The results of the listening test gave us information about the minimal audible angles for each plane. In the second study which was the main study, a localization test was performed, users had to localize sound sources in three conditions, there were several rounds in which the sources were separated by different angular spacing. The results of this study showed us if and how much the vertical plane impacts the localization performance and perception of a virtual audio environment.

5.1 Summary and contributions

We found out that the resolution of the HRTF-based Klang rendering is far better than other systems which do not use HRTFs and even surpasses those that do use general HRTFs. Simulating the elevation, i.e., including pitch and roll head movement angles in the rendering, did not significantly improve the recognition rate of sources in the horizontal plane. Furthermore there was no significant improvement in localization performance when sources were placed at different heights. We also did not discover any improvement in perception of the virtual audio space when elevation was simulated.
Since nowadays modern mobile devices are cheap, have high processing power and modern sensors are an integral part of these, we recommend using HRTF-based algorithms. Our findings show that there is an improvement of up to 50% in angular distance compared to other systems e.g. (Heller and Borchers [2015]). From the results of the different angular spacing used in the second experiment, we recommend using an angular spacing of about 15° at a distance of 2 meters to the user to be able to successfully discern between sources in the horizontal plane.

5.2 Future work

Some of the users in the second experiment informed us about missing depth, since we did not simulate room acoustics. We could additionally add reverb into the rendering to give a more natural impression of the room, this could further improve the localization performance and overall perception of the virtual audio space.
Furthermore the participants were all standing in a fixed position and only head movements were allowed, so only these were tracked and analyzed. To be even more realistic and reproduce a more natural orientation behavior for example in a museum, one could extend this study and investigate the impact of simulated elevation on navigation performance as in [Heller et al., 2014], i.e., the participants could be allowed to move around in the physical space so

that they can walk towards the sources and determine their position, the position of the participant would then have to be tracked additionally. One could for example analyze the paths which the subjects make in the two tasks and compare the distance efficiency.
Another interesting study would be to conduct an experiment similar to the one [Perrott and Saberi, 1990] did. One could investigate if increasing the elevation angle of a source reduces the horizontal minimal audible angle of two close-by placed sources. This could be done by virtually simulating a rotatable boom like in the former paper and additionally constructing a rotatable boom with speakers placed on it as a visual stimuli.

Appendix A

Presence Questionnaire

Localization-Test Questionnaire

User ID: Age: Sex:

Do you have a hearing disorder?	Yes	No
Do you have problem with spatial hearing?	Yes	No
Do you have experience with Augmented Audio Reality?	Yes	No

Run 1:

	Responsive		**Neutral**		**Not at all**
1. How responsive was the environment to actions that you performed?					
	Natural		**Neutral**		**Artificial**
2. How natural did your interactions with the environment seem?					
	Consistent		**Neutral**		**Not at all**
3. How much did your experience in the virtual environment seem consistent with your real-world experiences?					
	Very Well		**Neutral**		**Not at all**
4. How well could you localize sounds?					
	Quickly		**Neutral**		**Slowly**
5. How quickly did you adjust to the virtual environment experience?					

Figure A.1: The presence questionnaire which the participants had to fill out after each of the first two conditions in the second experiment. The questions were answered in a 5 points Likert scale with 1 being the best rating and 5 being the worst.

Bibliography

Paul Bertelson and Monique Radeau. Cross-modal bias and perceptual fusion with auditory-visual spatial discordance. *Perception & Psychophysics*, 29(6):578–584, 1981. ISSN 0031-5117. doi: 10.3758/BF03207374. URL http://dx.doi.org/10.3758/BF03207374.

Adelbert W. Bronkhorst. Localization of real and virtual sound sources. *The Journal of the Acoustical Society of America*, 98(5):2542–2553, 1995. doi: http://dx.doi.org/10.1121/1.413219. URL http://scitation.aip.org/content/asa/journal/jasa/98/5/10.1121/1.413219.

D. J. Kistler F. L. Wightman. Headphone simulation of free-field listening. ii: Psychophysical validation. *Acoustic origins of individual differences in sound localization behavior*, 85(2):858–878, 1989.

Florian Heller. Corona: Audio ar for historic sites. *AR[t] - Magazine about Augmented Reality, art and technology*, 5:80–85, May 2014. ISSN 2213-2481. URL http://arlab.nl/sites/default/files/ARt5_magazine_webversie.pdf.

Florian Heller and Jan Borchers. Audioscope: Smartphones as directional microphones in mobile audio augmented reality systems. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 949–952, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3145-6. doi: 10.1145/2702123.2702159. URL http://doi.acm.org/10.1145/2702123.2702159.

Florian Heller, Aaron Krämer, and Jan Borchers. Simpli-

fying orientation measurement for mobile audio augmented reality applications. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 615–624, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2473-1. doi: 10.1145/2556288.2557021. URL http://doi.acm.org/10.1145/2556288.2557021.

Simon Holland, David R Morse, and Henrik Gedenryd. Audiogps: Spatial audio navigation with a minimal attention interface. *Personal and Ubiquitous Computing*, 6(4): 253–259, 2002.

Torben Jørgensen Jan Abildgaard Pedersen. Localization performance of real and virtual sound sources. In *Proceedings of the NATO RTO-MP-HFM-123 New Directions for Improving Audio Effectiveness Conference*, pages 29–1 to 29–30. NORTH ATLANTIC TREATY ORGANISATION, 2005.

T. Sadralodabai D. R. Perrott K. Saberi, L. Dostal. Minimum audible angles for horizontal, vertical, and oblique orientations: Lateral and dorsal planes.

Nicholas Mariette. Navigation performance effects of render method and head-turn latency in mobile audio augmented reality. In Sølvi Ystad, Mitsuko Aramaki, Richard Kronland-Martinet, and Kristoffer Jensen, editors, *Auditory Display*, volume 5954 of *Lecture Notes in Computer Science*, pages 239–265. Springer Berlin Heidelberg, 2010. ISBN 978-3-642-12438-9. doi: 10.1007/978-3-642-12439-6_13. URL http://dx.doi.org/10.1007/978-3-642-12439-6_13.

John C. Middlebrooks. Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency. *The Journal of the Acoustical Society of America*, 106(3):1493–1510, 1999. doi: http://dx.doi.org/10.1121/1.427147. URL http://scitation.aip.org/content/asa/journal/jasa/106/3/10.1121/1.427147.

David R. Perrott and Kourosh Saberi. Minimum audible angle thresholds for sources varying in both elevation and azimuth. *The Journal of the*

Acoustical Society of America, 87(4):1728–1731, 1990. doi: http://dx.doi.org/10.1121/1.399421. URL http://scitation.aip.org/content/asa/journal/jasa/87/4/10.1121/1.399421.

Ville Pulkki. Virtual sound source positioning using vector base amplitude panning. *J. Audio Eng. Soc*, 45(6):456–466, 1997. URL http://www.aes.org/e-lib/browse.cfm?elib=7853.

Ville Pulkki and Matti Karjalainen. Localization of amplitude-panned virtual sources i: stereophonic panning. *Journal of the Audio Engineering Society*, 49(9):739–752, 2001.

M. A. Ericson R. L. McKinley. *Binaural and spatial hearing in real and virtual environments*, chapter Flight demonstration of a 3-D auditory display, pages 683–699. Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc, 1997.

Christian Sander, Frank Wefers, and Dieter Leckschat. Scalable binaural synthesis on mobile devices. In *Audio Engineering Society Convention 133*, Oct 2012. URL http://www.aes.org/e-lib/browse.cfm?elib=16525.

Lucia Terrenghi and Andreas Zimmermann. Tailored audio augmented environments for museums. In *Proceedings of the 9th International Conference on Intelligent User Interfaces*, IUI '04, pages 334–336, New York, NY, USA, 2004. ACM. ISBN 1-58113-815-6. doi: 10.1145/964442.964523. URL http://doi.acm.org/10.1145/964442.964523.

Elizabeth Wenzel, Frederic Wightman, Doris Kistler, and Scott Foster. Acoustic origins of individual differences in sound localization behavior. *The Journal of the Acoustical Society of America*, 84(S1):S79–S79, 1988. doi: http://dx.doi.org/10.1121/1.2026486. URL http://scitation.aip.org/content/asa/journal/jasa/84/S1/10.1121/1.2026486.

Elizabeth M. Wenzel, Frederic L. Wightman, and Doris J. Kistler. Localization with non-individualized virtual acoustic display cues. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '91, pages 351–359, New York, NY, USA, 1991. ACM. ISBN

0-89791-383-3. doi: 10.1145/108844.108941. URL http://doi.acm.org/10.1145/108844.108941.

S. Werner, J. Liebetrau, and T. Sporer. Audio-visual discrepancy and the influence on vertical sound source localization. In *Quality of Multimedia Experience (QoMEX), 2012 Fourth International Workshop on*, pages 133–139, July 2012. doi: 10.1109/QoMEX.2012.6263855.

György Wersényi. Localization in a hrtf-based minimum-audible-angle listening test for guib applications. *Electronic Journal "Technical Acoustics"*, 2007.

G. B. Wetherill and H. Levitt. Sequential estimation of points on a psychometric function. *British Journal of Mathematical and Statistical Psychology*, 18(1):1–10, 1965. ISSN 2044-8317. doi: 10.1111/j.2044-8317.1965.tb00689.x. URL http://dx.doi.org/10.1111/j.2044-8317.1965.tb00689.x.

Mengqiu Zhang, Wen Zhang, Rodney A Kennedy, and Thushara D Abhayapala. Hrtf measurement on kemar manikin. *Proc. ACOUSTICS'09 (Australian Acoustical Society)*, page 8, 2009.

Index

Typeset October 15, 2015