RWTH AACHEN UNIVERSITY

Orientation Measurement for Mobile Audio Augmented Reality Applications

Bachelor's Thesis at the
Media Computing Group
Prof. Dr. Jan Borchers
Computer Science Department
RWTH Aachen University

by
Aaron Krämer

Thesis advisor:
Prof. Dr. Jan Borchers

Second examiner:
Prof. Dr. Torsten Kuhlen

Registration date: 5.11.2013
Submission date: 2.04.2014

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, April 2014
Aaron Krämer

Contents

List of Figures

Abstract

Audio-augmented reality is a method to augment the environment and objects in the real world with virtual sounds in a given context. Therefore, position and orientation of the user has to be tracked. Most of the currently existing mobile applications are tracking the head orientation. Since users are mobile, the easiest way to provide audio to the user would be using headphones. And while one are using headphones, the idea of tracking orientation at the head is not far away. To do so, they have to mount additional hardware on the user which could be easier with headphones, since they have an own frame. We are going to investigate if we can track the orientation at a different location than the head, e.g., at the body or device, which the user holds in her hand. This could reduce the effort of mounting extra hardware on the user which already in current smart phones.
We conducted two studies. The first one was lab-based and should show us how users orient towards sound sources. The second one was conducted in the Coronation Hall in the historic city hall of Aachen to figure out the users perception of different tracking positions.

Überblick

Audio-augmented reality ist eine Methode zur Erweiterung unserer Umgebung und Objekten der realen Welt mit virtuellen Sounds in Bezug auf den jeweiligen Kontext. Dafür muss die Position und Orientierung des Benutzers verfolgt werden. Bisher haben die meisten mobilen Systeme immer die Kopfbewegung verfolgt. Da die Benutzer sich bewegen ist wäre der einfachste Weg Kopfhörer zu benutzen um die Geräusche wiederzugeben. Und wenn schon Kopfhörer benutzt werden liegt die Idee die Orientierung des Kopfes zu erfassen nicht fern. Um das zu bewerkstelligen musste bisher immer extra Hardware am Benutzer befestigt werden welche am Kopfhörer gegebenenfalls leichter anzubringen ist da sie bereits ein eigenes Gerüst besitzen. Wir möchten mit dieser Arbeit herausfinden ob die Orientierung auch an einer anderen Stelle wie z.B. dem Oberkörper oder der Hand, in welcher ein Gerät gehalten wird, gemessen werden kann. Dadurch könnte der Aufwand verringert werden extra Hardware am Nutzer zu befestigen welche eh schon in heutigen Smartphones vorhanden ist.

Aus diesem Grund habe ich zwei Studien durchgeführt. Die erste Studie wurde unter Laborbedingungen durchgeführt und sollte uns zeigen wie Benutzer sich zu Soundquellen orientieren. Die zweite Studie wurde im Krönungssaal des Aachener Rathauses durchgeführt um herauszufinden wie die Benutzer die unterschiedlichen Positionen von Orientierungsmessungen empfinden.

Acknowledgements

I would like to thank Kerstin, Ben, and Marcel for helping me with the correction and Flo for his persistent help! Also, I like to thank all the users who participated in 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.

Chapter 1

Introduction

An audio-augmented reality installations extends the real world with virtual sounds. These sounds could be context dependent, e.g., frog sounds that are coming from a lagoon or they could be unrelated, e.g., animal sounds that are coming from stone statues which do not represent animals [Vazquez-Alvarez et al., 2012]. The audio, that comes from the virtual sound sources is altered depending on the users position and orientation. The immersion is quite perfect so that the user gets the impression that the sounds were emitted from the real world.

Most applications track the head.

Most of the current existing mobile audio augmented reality installations are tracking the head orientation of a user. Since users are mobile, the easiest way to provide audio to the user would be using headphones. And while one is using headphones, the idea of tracking orientation at the head is not far away. To do so, one has to mount additional hardware on the user which could be easier with headphones since they already have an own frame. While using headphones with the equipped hardware, rotation of the head describes the orientation of the user. In the past, most installations were using complex and large hardware. For example, [Holland et al., 2002] used a notebook in a rucksack running the software, equipped with additional external hardware like GPS transceiver and compass modules.

Additional hardware is needed.

So, why not use smartphones, which users already have with them. They are already equipped with the needed hardware like a GPS transceiver and electronic compass, and have at least enough power to compute spatial audio rendering. One possible problem of smartphones would be the position of the tracking as users are having their device in their pocket or hand.

We want to figure out if we can track orientation at other locations.

Since most common installations, as already mentioned, are tracking the users head, we have to figure out if a tracking location other than the head is possible. We will investigate whether it is possible to track the orientation of the user at different positions, as shown in Figure 1.1, without reducing the immersion for the user of the virtual sound space. As we are more interested in the field of audio guides, where users are holding the device in their hand, we do not investigate the position of the pocket.

We conducted two studies to answer the following questions:

- How do users orient towards sound sources?
- Is it possible to use tracking of the body or device instead of the head to orient in virtual sound installations without any perceived difference?

In the following we first talk about other existing applications and results in the field of audio-augmented reality. Then we describe the implementation of our own audio-augmented reality application and technical setup of the experiments. Afterwards we present the conditions and methodologies of our two experiments just as the results and a discussion. In the end, we give a summary of this thesis and an outlook of future studies.

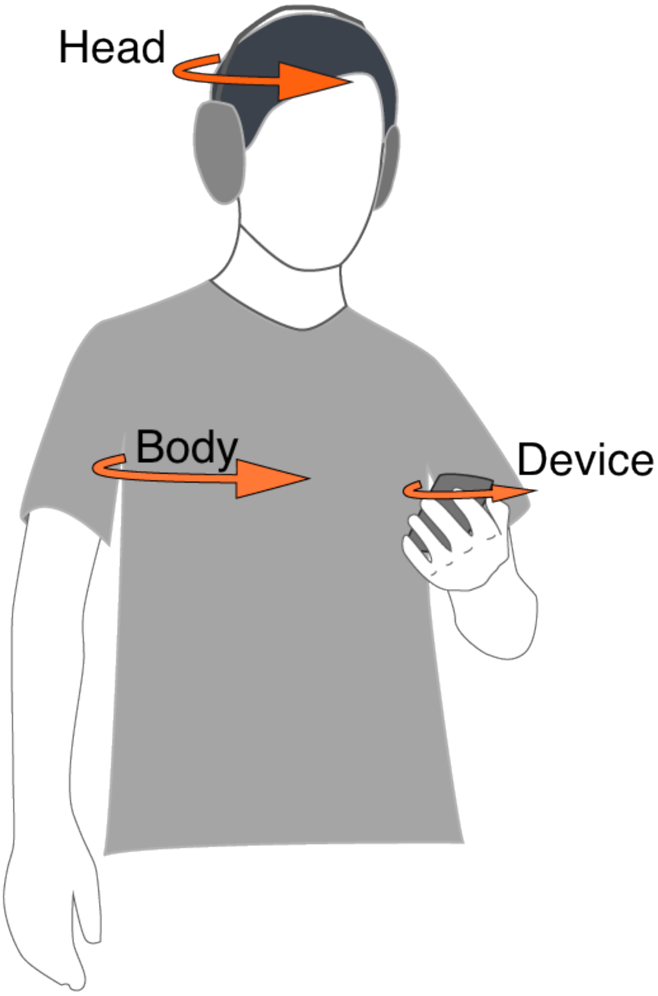


Figure 1.1: The three compass sensor placements we are investigating on.

Chapter 2

Related Work

One of the first field tests with audio-augmented reality as navigational aid was conducted by [Holland et al., 2002]. Their prototype, called AudioGPS, is a spatial audio user interface. They analyzed various audio mappings to represent location and direction. All sounds are non-speech and non-continuous because they want to avoid additional load on the human voice channel. They argue that speech sounds will place a large processing and attention burden on the user. To provide direction to the user, they use simple stereo panning to move an audio source around the users head. Two different sounds are used to distinguish between the semicircle in front and semicircle behind the user. Since the hardware did not include a compass, the direction of motion is calculated when the user is moving. To provide distance, they use a principle similar to the 'Geiger counter' metaphor . This means that the number of pulses of sound together with their rapidity gives an indication of how far away a given point is.
Their prototype did not use an electronic compass to get the direction of the user. Therefore they have a latency of 10 to 15 seconds before the system starts reporting an update.

Various audio mappings were analyzed to represent location and direction to the user.

The 'Geiger counter' metaphor was used to provide distance information.

Another audio-augmented reality navigation application is the roaring navigator [Stahl, 2007], a group guide for a zoo with a shared auditory landmark display. An auditory display uses non-speech sounds to present information. This one is related to landmarks too, specifically to the sound of

Group guide for a zoo using auditory landmarks.

animals around the zoo.

People are going in groups of two (master and slave device) along a path or can just explore the area on their own. During that time, both share the same audio playback of the master device, but every user's audio is altered depending on their relative position to the virtual sound position. Distance is provided by altering the volume in both ears, direction by altering the volume on the left or right side of the headset (simple audio panning). A magnetometer was clipped on the back of a baseball cap that users were wearing during the experiment to get the orientation.

A virtual sound garden using Earcons to identify landmarks.

Similar to the roaring navigator, [Vazquez-Alvarez et al., 2012] built up a virtual sound garden placed in a park in Funchal, Madeira. They placed Earcons at specific positions of landmarks of this park.

Definition: *Earcon*

EARCON:
An Earcon is a non-verbal audio message which uses an abstract mapping to provide information to the user [Vazquez-Alvarez et al., 2012].

Although, the mapping of the Earcons is abstract and not related to the landmarks they were placed on, they used them to identify the landmarks. They thought that the Earcons fit well in the context of the park since these Earcons had the sounds of typical animals. Each landmark was software wise surrounded by an activation (10 m radius) and proximity zone (25 m radius). If the user enters the proximity zone the appropriate Earcon gets played to signalize the user that a landmark is close. By entering the activation zone, the user has the possibility to press a button on the device to start an audio clip with information about the sites.

A Nokia N95-8GB connected to a GPS receiver and a JAKE sensor pack was used to run the application and track the position and heading. Four different rendering features were implemented: *Baseline* (no Earcons or audio spatialization), *Earcons* (no audio spatialization), *Spatial* (basic proximity zone with Earcons and limited audio spatialization (distance)), and *Spatial3D* (Earcons and audio spatialization).

3D spatial audio rendering together with Earcons was the most effective technique as their results show. Although, users spent more time and walked a greater distance when audio spatialization was used, it did not lead to frustration, rather it leads to a greater enjoyment and discovery of the participants.

Users took longer and walked greater distances, but enjoyed the experience more.

[Ankolekar et al., 2013] analyzed the performance and emotional engagement of different types of audio-based clues for directing users' attention like the Earcons mentioned above. Users were interrupted by audio clues while walking on a shopping street. The audio clues were played for a minute and the users had then to identify on a map which POI (Point Of Interest) was meant. For participants in the treatment group each of these five different cues was played: visual cue, speech cue, auditory icon (representative sound of a place), musicon (fragment of music that could be representative for that place), and a *Mix* condition (plays first the auditory icon, then the speech cue and at last the musicon for the remainder of the minute). The control group only received visual clues. Their results show that musicons would be the better choice for serendipitous discovery, pleasure and identification accuracy.

Different audio clues were used to to analyze performance and emotional engagement while directing users.

[Marentakis and Brewster, 2006] studies the field of spatial audio displays. In detail they are interested in pointing to virtual spatial audio sources. Contrary to the other approaches mentioned before, they follow an egocentric approach instead of an exocentric . This means that the sound position is fixed to the user, independent of the users direction. They hypothesize that target width and distance to target are affected in a manner similar to Fitts' law. In their study, users were separated into two groups; with target feedback and without. Every group had to perform pointing tasks towards sounds while standing and walking. Therefore, another question is how mobility affects selection times and selection accuracy.

Spatial target acquisition with an egocentric approach.

Their results showed a significant effect for mobility and feedback. Mobility leads to slower and less accurate interaction, where feedback decreases the interaction speed but increase the interaction accuracy. Additionally, Fitts' Law could be applied for specific target widths.

Mobility and feedback influences walking speed, interaction accuracy and speed.

Analysis of head-turn latency in mobile audio-augmented reality.

One of the main parts of Mariette's work [Mariette, 2010] is the analysis of head-turn latency in mobile audio-augmented reality. In his study users had to walk towards virtual sound sources, placed on the frame of a circle. Every user was equipped with a handheld computer, interfaced with position and orientation tracker. Users had to start at the middle of the circle for each trial. All users did several trials to different source positions. He encouraged them to look ahead while they walked, and use head-turns to find the correct source direction. If the user failed to locate the source within 60 seconds, the source stops playing and a time out message was displayed on the screen.

In addition he tested the system latency. He came to the conclusion that head-turn latencies up to 176 ms and total system latencies up to 376 ms can be tolerated until effects can be observed regardless of the rendering method.

Chapter 3

Setup

Many mobile audio-augmented reality applications are using the head as the source of orientation of the user. This means that these systems alter the audio related to the movement and rotation of the head. To get this information, additional sensors are required, which need to be placed onto the head and connected in hard- and software. Current smartphones already have the desired hardware on board. If we can use these sensors to create a compelling experience, this would reduce the effort of implementation. Therefore we want to figure out if we can track the orientation at the position of the hand or body. We are more interested in the position of the hand. Since our scenario is a museum guide, users are already holding their devices in front of them in their hands.

Can we track the orientation at a different location to reduce implementation effort?

We conducted two studies. With our first lab-based study we wanted to figure out how users orient towards sound sources. Therefore we let users perform some orientation tasks and tracked the orientation of head, body, and device. In the second study we were interested in the perception of the user if we track the orientation at different positions. We let users walk three different paths with different tracking positions through the Coronation Hall in the historical city hall of Aachen (Figure 3.1).

Two studies were conducted.

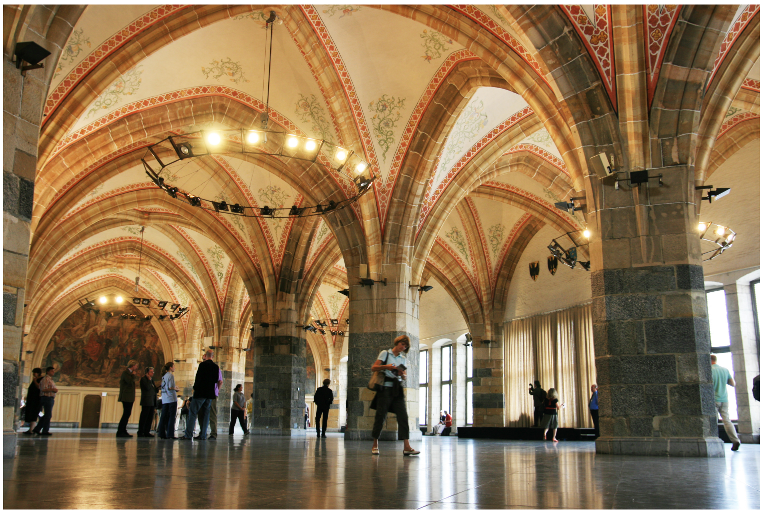


Figure 3.1: Coronation Hall in the historical city hall of Aachen.

3.1 The Corona Audio Space

We are using *Corona* during the experiments therefore we will describe it now before we go over to the implementation of *Corona* and technical setup of the experiments.

Conversations of a medieval ceremony placed virtually in the Coronation Hall.

Corona [Heller et al., 2009] is an audio-augmented reality installation in the historical city hall of Aachen. As a visitor you are wearing headphones and holding a museum guide in your hands while walking through the Coronation Hall (Figure 3.1). *Corona* takes the visitor into a medieval ceremony; the coronation of Charles V. Conversations of people of this medieval ceremony, who are talking about different aspects of the ceremony, are placed virtually in the Coronation Hall. The visitor does not see any related objects at this position in the real world. While you are walking through the Coronation Hall, the audio is panned around your head depending on your orientation and the volume is altered depending on your distance to the target.

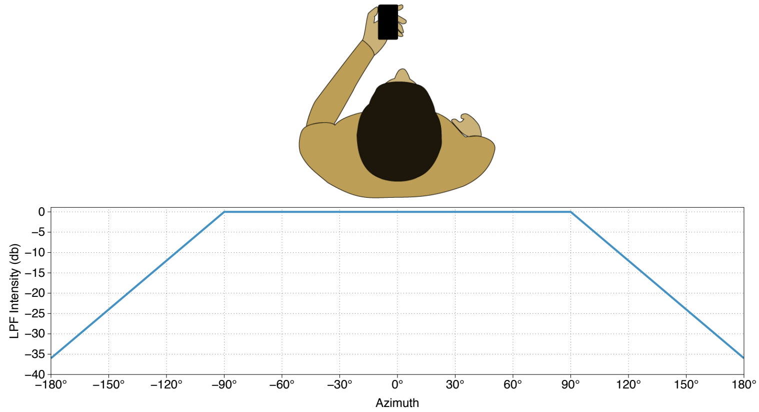


Figure 3.2: A low-pass filter is applied to sources in the back to reduce front-back confusions in our spatial audio rendering.

3.2 Implementation

The audio rendering is performed on an Apple iPhone 4S running iOS 5.1.1 using the OpenAL Framework with the `ALC_EXT_MAC_OSX` extension. This extension provides a more realistic spatialization based on a spherical head model and including the following filter factors: interaural level difference, interaural time difference, head filtering, and frequency dependent distance filtering. We used the `ALC_EXT_ASA` extension to improve the perception of sources that are behind the user. This extension enables additional effects such as reverb, obstruction, and occlusion. To overcome front-back confusion , which is a common problem in virtual sound spaces [Bronkhorst, 1995], a low-pass filter was added that muffles the sound behind the user. The low-pass filter intensity is interpolated linearly between 0 dB and 36 dB for sources with an azimuth angle between 90° and 180° (Figure 3.2).

Detailed description of the implemented audio-augmented reality application.

We modified the implementation of *Corona* for our experiments. Sound files were replaced and the graphical user interface was adjusted to our needs. We also tuned the audio rendering parameters to fit our room and experiment conditions. The whole implementation was adopted from [Heller et al., 2014].

3.3 Technical Setup: First Experiment

The goal of the first experiment was to figure out how users orient towards virtual sound sources. We build up a circle with 4 m diameter and placed 24 Wavemaster Mobi loudspeakers in a height of 140 cm and an angle of 15° between each of them (Figure 1.1). We created the circle with sound sources to have a reference to the work of [Mariette, 2010]. The headphone we used was an AKG K-512.
The user was equipped with markers on the head, at the waist and the device which she was holding in her hand.

Vicon optical tracking system.

As we used a Vicon optical tracking system, the headphones were equipped with optical tracking markers. In the condition without headphones, users had to wear a headband with markers. All markers were tracked simultaneously. The positions of the markers were transported from the Vicon system to our application via WiFi. The update rate of the Vicon is 100 Hz. The Vicon tracker has a latency of 2.5 ms and the average round trip time of the WiFi connection was 4.7 ms. So we are below the total system latency of 376 ms mentioned by [Mariette, 2010].

3.4 Technical Setup: Second Experiment

Our second experiment takes place in the Coronation Hall in the historical city hall of Aachen. There, we placed six virtual sound sources in form of a rectangle (Figure 3.4). One on every corner and the remaining two are placed in the middle of the long edge.
The Coronation Hall is too large (20 × 45 m) for an optical tracking system, so we used the Ubisense RTLS [1] tracking system.

Ubisense location tracking system.

Ubisense is a real time location tracking system that works with tags and receivers. Ubisense-tags send out ultra-wideband pulses which are received by the antennas of the Ubisense sensors. The Ubisense system has an accuracy of 15 cm in the center of the covered area and 50 cm at the outer borders. The refresh rate is approximately 10 Hz. The location measurement has a specified latency of 234 ms

[1] http://www.ubisense.net

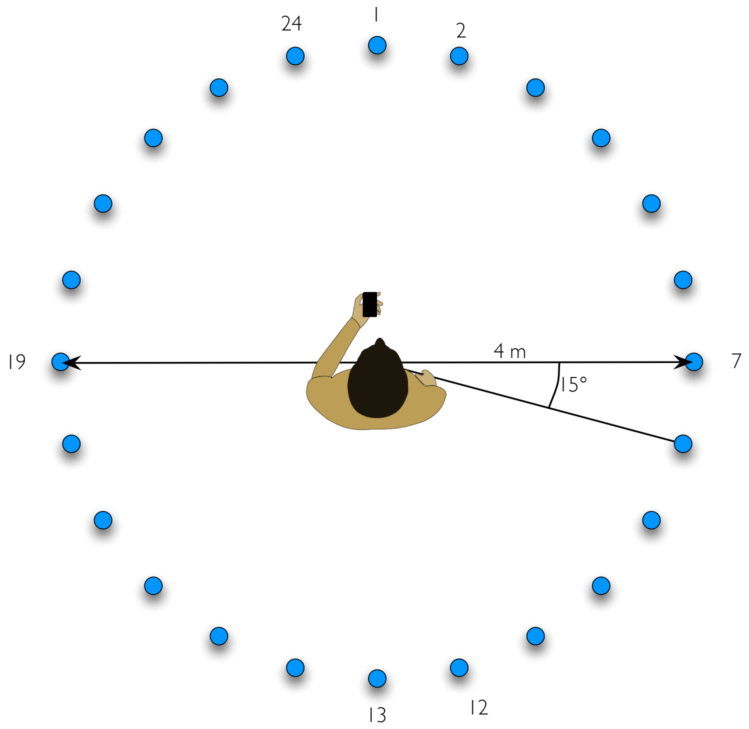


Figure 3.3: Technical setup of the first experiment.

and the transport of the data through WiFi takes 42 ms in average. The approximate overall delay of the system is then 276 ms. This is below the total system latency of 376 ms mentioned by [Mariette, 2010]. To measure the orientation of the user we used an external tilt-compensated compass (HMC6343) with a refresh rate of 10 Hz.

Total system latency of 276 ms.

[Walker and Lindsay, 2006] and [Mariette, 2010] concluded that a capture radius between 1.5 meters and 2 meters would be best. This capture radius is a threshold to signalize the software that a user has reached a desired sound source, if the user is within the radius. Because of the delay of the tracking system and due to some trials of our own, we decided to use a capture radius of 2 meters.

Capture radius and activation zone.

The sound sources are automatically played if the user reaches the activation zone of a sound source and paused while leaving the activation zone. The sound source placed along the long side of the Coronation Hall have a distance to each other of 14 meters. The distance between the sources along the short edge is 8 meters. Through some testing we found out that 9 meters fits best as radius for the activation zone to let the user hear at least one other sound source to have a clue of another sound source and to reduce the volume of other surrounding sources to a minimum to not disturb the user.

Our implementation of capture radius and activation zones are similar to the activation and proximity zones used by [Vazquez-Alvarez et al., 2012].

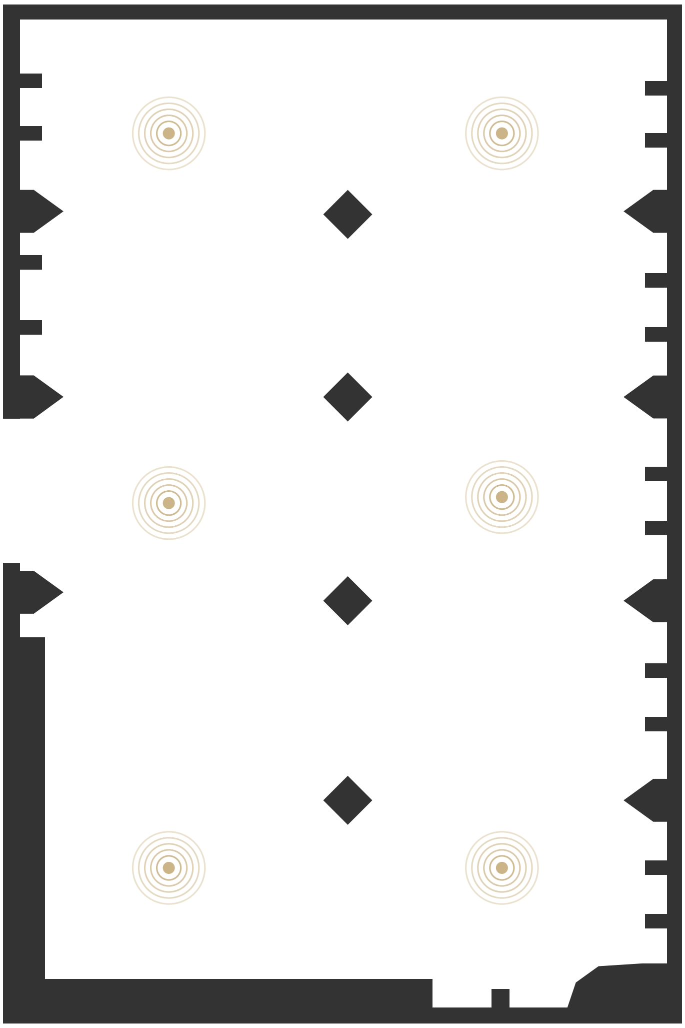


Figure 3.4: Source placement in the Coronation Hall for our second experiment. The free area on the left is the entrance. The yellow circles show the source positions.

Chapter 4

Evaluation

4.1 First Experiment: Movement and Orientation

In this first experiment we wanted to investigate how users move and orient towards sound sources. Since we are interested in the head turns the users did, the room where this study took place was very small (approximately 5 × 5 m). Users do not had to go long paths to the sound sources.

4.1.1 Conditions and Methodology

Experiments show that there are performance differences between speech and non-speech sounds [Tran et al., 2000]. Therefore, we used two different sound types. A monologue, spoken by a male voice is used as speech sound and a drum sound was used as the non-speech sound. We used two different rendering techniques; headphone and speaker. Loudspeakers simulate perfect spatial audio rendering therefore we used them as one condition to control our results since we did a within-subjects study. In the headphone condition we used our audio-augmented reality application with spatial audio rendering. The users had to wear headphones. For the speaker condition we just played the sounds through the loudspeakers. For ev-

Two rendering and two sound-type conditions.

ery rendering, users had to perform tasks with speech and non-speech sounds. The virtual sound sources are placed at the real positions of the loudspeakers (Figure 1.1). During the speaker condition a cable was hanging from the ceiling down to the center of the circle which was connected to the device the user was holding. It is used to play the sounds through the loudspeakers.

24 users, 3 female, in the age 19 - 53 (average 26) participated in this study. Every user had to perform 12 learning trials with headphones, six with speech and six with non-speech sounds, followed by 24 trials per condition; in sum 108 trials.

Description of trials the user had to do.

The user started each trial in the middle of the circle (Figure 1.1) holding the device in her hands and facing into the direction of Source 1. To start one trial a button on the device had to be pressed, thereafter the sound had to be located. Another button was pressed in front of the sound source when the user thought that the right one was located. Then she had to go back to the middle and repeat the procedure for every source. The order of the occurring sources was randomized for all conditions. Due to software failures we had to remove sound source 7. Therefore, 23 of the 24 sound sources are left to be analyzed.

4.1.2 Results

Following the definition of [Mariette, 2010], head-yaw (θ_h) is the relative angle of the head to the body. Device-yaw (θ_d) is defined as the relative angle between device and body. Head-device-yaw (θ_{hd}) is the relative angle between head and device. We transformed the values from their reported range of $0°$ to $360°$ to $[-180, ..., 180]°$, with $0°$ being the direction of the user's torso. We subtracted the initial difference between head, device, and body at the beginning of each trial since this difference is assumed to be caused by the placement of the tracking markers.

Transformation of the reported values.

The calculated mean for θ_h with headphone rendering is $M = -1.57°, SD = 15.83$ and $M = -2.24°, SD = 19.98$ for speaker. Head and body are aligned most of the time as the means show. The means for θ_d are $M = -0.17°, SD = 8.62$ for headphone and for speaker $M = -0.35, SD = 13.05$.

Body and device are aligned more than head and device.

In the calculation of means, positive and negative angles will cancel each other out. Therefore we calculated the root mean square (RMS) of head-yaw ($\theta_{h(\text{RMS})}$) and device-yaw deviation ($\theta_{d(\text{RMS})}$). On these average values of head and device turns we did a log-transform and performed a repeated measures ANOVA. This shows a major effect of the used rendering on $\theta_{h(\text{RMS})}$ ($F(1, 2106) = 111.17, p < 0.0001$). The RMS means differs only by $4°$ (Headphone: $M_{\text{RMS}} = 13.86°, SD = 8.05$ and Speaker: $M_{\text{RMS}} = 17.75°, SD = 9.93$). If we compare the RMS means angle between head and device $\theta_{hd(\text{RMS})}$ (Headphone: $M_{\text{RMS}} = 15.06°, SD = 9.19$, Speaker: $M_{\text{RMS}} = 19.73°, SD = 11.84$) and head and body $\theta_{h(\text{RMS})}$ we see that they are in the same range. It shows that body and device orientation could be assumed as equal in this case.

Body and device orientation could be assumed as equal.

The source position also has a major effect on $\theta_{h(\text{RMS})}$ ($F(23, 2086) = 21.48, p < 0.0001$). When users orient towards sound sources behind them, they naturally do larger head-turns.

4.1.3 Discussion

Since we presented information on the screen the user had to interact with, she will probably hold the device in front of her body. In the usage scenario of *Corona* users also hold their device in front of them, as it shows them additional information to exhibits on the screen. This will also explain the smaller θ_d values compared to θ_h.

Users were doing a large initial head-turn to get the initial orientation (Figure 4.2). In both renderings this distinctive head turn is present and looks very similar. After that initial head-turn, head-yaw seems to stay nearly within a $10°$ angle to both sides. If we would track body or device, we would risk loosing the large initial head-turns in the beginning. This could degrade the presence of the virtual sound space for the users as they are used to get a first orientation. Figure 4.3 shows the initial head turn of all users in all conditions. The mean duration time of the peaks exceeding $15°$ in the first 4 s of each trial is 590 ms. Again, if we are going

Users did large initial head-turns to get the initial orientation.

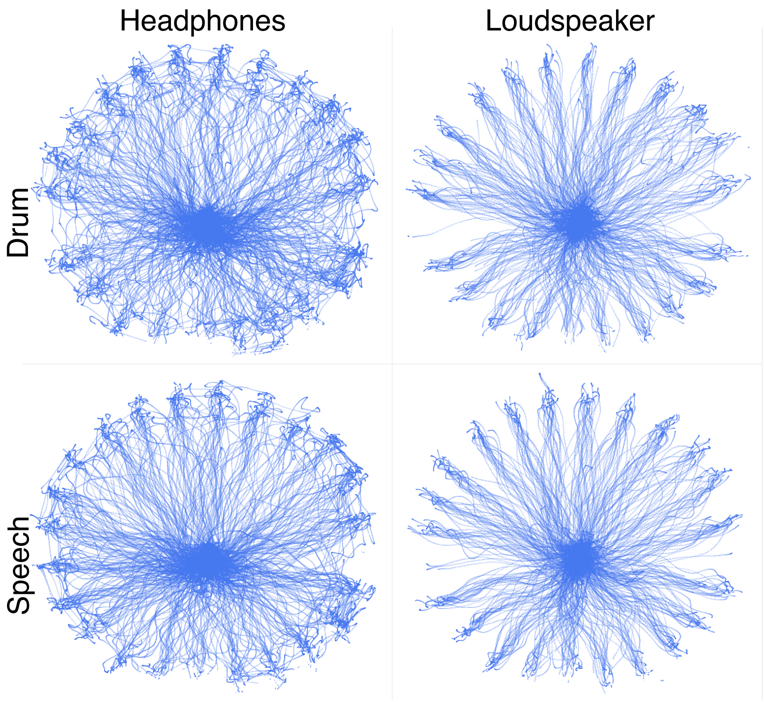


Figure 4.1: Paths of participants from the start in the center to sources on the border of the circle.

Using different tracking locations than the head, we risk loosing these initial head-turns.

to use body or device tracking we may loose these 590 ms in the beginning and the user could perceive this as a delay. Mariette [2010] results show that these 590 ms are to big as not be recognized by the user, but completing navigational tasks would still be possible.

4.2 Learning Effects

Analysis of the learning tasks each user did before the experiment.

In the first experiment we let users perform 12 tasks with headphone to become familiar with our system. Six with the monologue speech sound and six with the non-speech drum sound. Conditions and methodology are the same as in the first experiment except the order of the occurring sound sources was equal for each user. So each user went the same path during the learning phase.
We tried to find learning effects by analyzing the data we

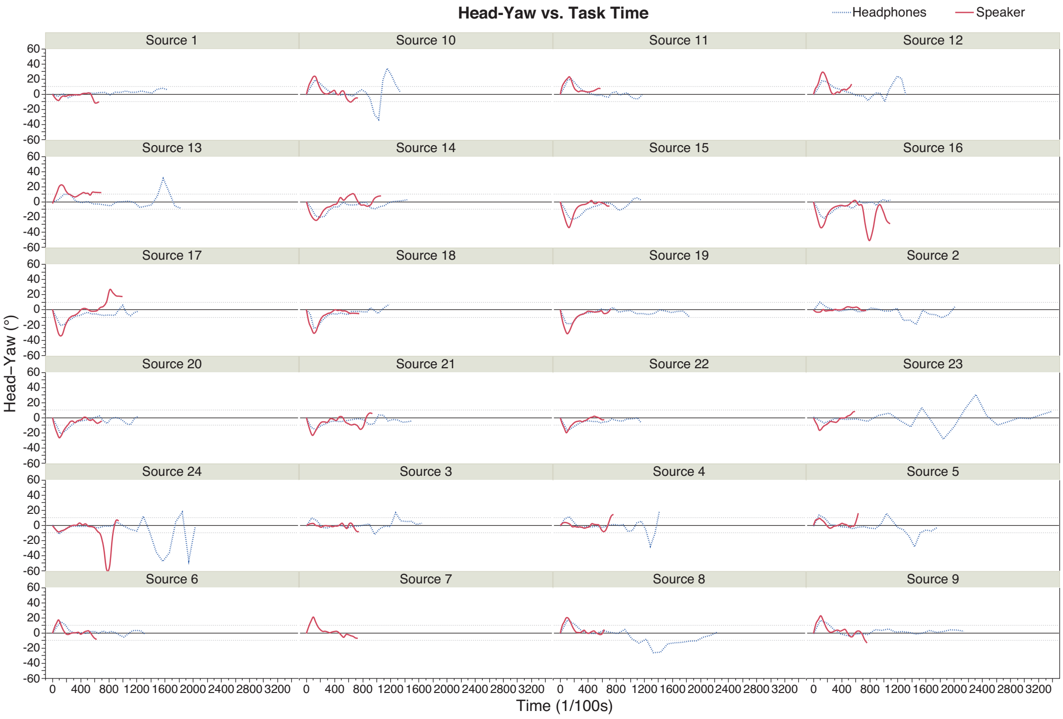


Figure 4.2: Mean head-yaw per source.

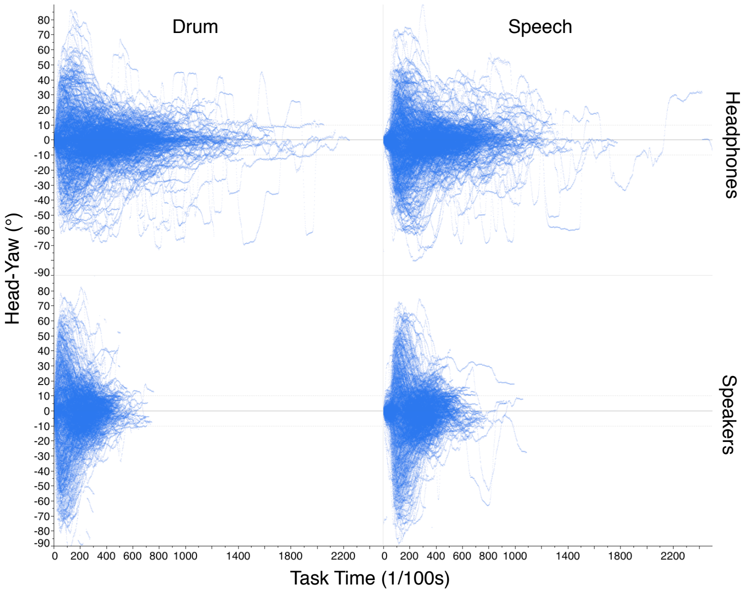


Figure 4.3: Head-Yaw over task time of all users in all conditions.

got from each user. For the analysis we looked at the time the user needed for every task. In the following we will present the results of the analysis.

4.2.1 Results

We had to remove the reported data of one user and as already mentioned in the first experiment we had to remove sound source 7 due to a software failure. Also, some tasks were not logged during the learning phase.

We calculated the mean time by user and task. Then we log-transformed the data and tested each task for a normal distribution. Except for task 4 ($W = 0.91, p = 0.0428$) and task 9 ($W = 0.89, p = 0.0256$), all tasks are normally distributed with task 12 ($W = 0.95, p = 0.4$) having the lowest significant p-value. A low p-value rejects the hypothesis that the data is normal distributed. Nonetheless, we ran an analysis of variance with task as effect model that showed us a significant difference ($(F(10, 3.0515), p = 0.0012$) between the positions of the tasks.

Specific source positions differ in time from each other.

We did a pairwise Student's t-test to find the significant tasks that differ. Task 5, which represents source 1 (Figure 3.3) significantly differs from source 22 ($p = 0.0001$), source 13 ($p = 0.0002$), source 19 ($p = 0.0003$), and source 14 ($p = 0.0064$). Also source 22 differs significantly from source 18 ($p = 0.0059$) and 23 ($p = 0.0092$).

4.2.2 Discussion

By looking at the order of sound sources the user had to walk to in the learning phase, we see that participants had to walk to source 13 in task 4 and in the next task to source 1. Source 13 is exactly on the opposite side of source 1 if we look again at the circle of sound placements (Figure 3.3). Since we added a low-pass filter for sounds behind the user, sounds in the back get muffled and therefore differ from sounds in the front. When sounds in front of the user get played the volume in both ears is equal. But there is no additional information provided to the user that, like muffling the sound, shows her that the source is in front of her. The user could be confused as she walked to the source in the back before and then to the source in the front and again, she hears the sound aligned in both ears. So we came up with the idea to test our data against front-back confusion and also against left-right confusion. Front-back confusion will be analyzed and explained in the next section.

Idea of testing against front-back confusion.

4.3 Front-Back Confusion

Front-back confusion is the phenomenon where a sound source is placed in the back and is perceived as being in the front or vice-versa.

We used the data of the first experiment and calculated the mean-time the user has taken per task. Then we grouped our 24 sound sources into four orientations: front, back, left, and right (Figure 4.4). Every group consists of the six sound sources which are in a 90° field-of-view of the users position depending on the mentioned orientation.

Sound sources were grouped in directions front, back, left, and right.

4.3.1 Results

We did an analysis of variance over the mean-time per orientation grouped by rendering and sound type. Except for rendering *Speaker* and sound type *Speech* ($F(3, 548) = 1.2247, p = 0.3$) we have significant differences. For rendering *Headphones* with sound *Non-Speech* ($F(3, 497) = 7.5916, p < 0.0001$), and with sound *Speech* ($F(3, 524) = 3.9802, p = 0.008$) we have a significant difference between *front* and *back* sources. For rendering *Speaker* and sound *Non-Speech* ($F(3, 548) = 3.1793, p = 0.0237$) we do not have such a high p-value as in the headphone condition.

We will now look at them in detail to see differences between single orientations. Therefore, we did a Student's t-test analysis by rendering and sound. Between orientations *front* and *back* we have significant differences for rendering *Headphones* with sound *Non-Speech* ($p < .0001$), and *Speech* ($p = 0.0056$) and for rendering *Speaker* with sound *Non-Speech* ($p = 0.0043$). For rendering *Speaker* with sound

Front and back directions significantly differ in nearly all conditions.

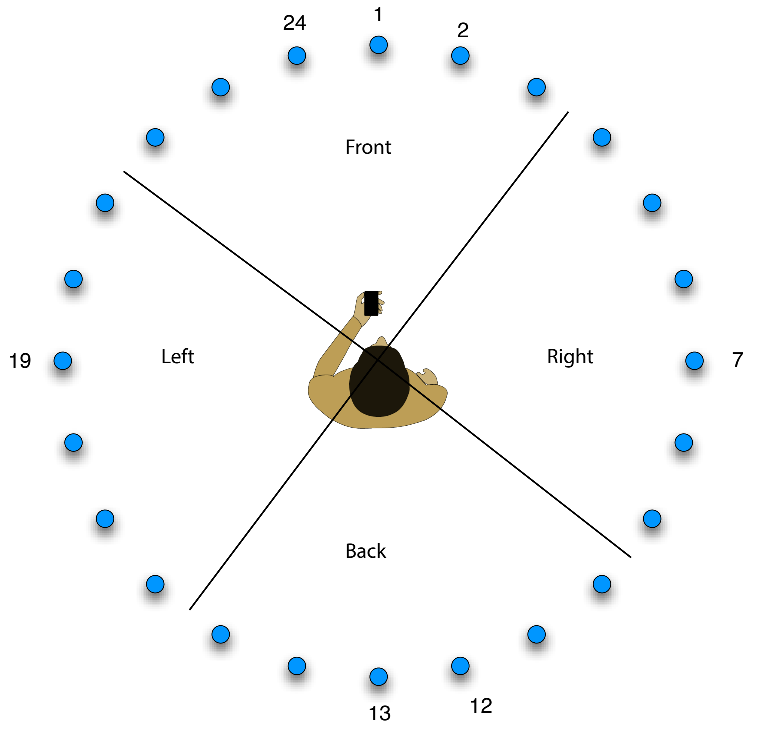


Figure 4.4: The four groups of direction to check against front-back and left-right confusion.

Speech we do not have any significant difference between the group of sources in the front and in the back.

4.3.2 Discussion

Rate of front-back confusion differs per condition.

In the headphone condition users took significantly longer to orient to the front than to the back for both sound-types (Figure 4.5). In the speaker condition only with sound-type non-speech we have a significant difference, but not as high as in the headphone condition. So we have definitely a higher rate of front-back confusion in the scenario with virtual sound sources as in the scenario with real sound sources. These results fit perfectly with the results of [Bronkhorst, 1995]. They showed that the rate of front-back confusions is higher with virtual sound sources than with real sound sources. The results of our analysis does

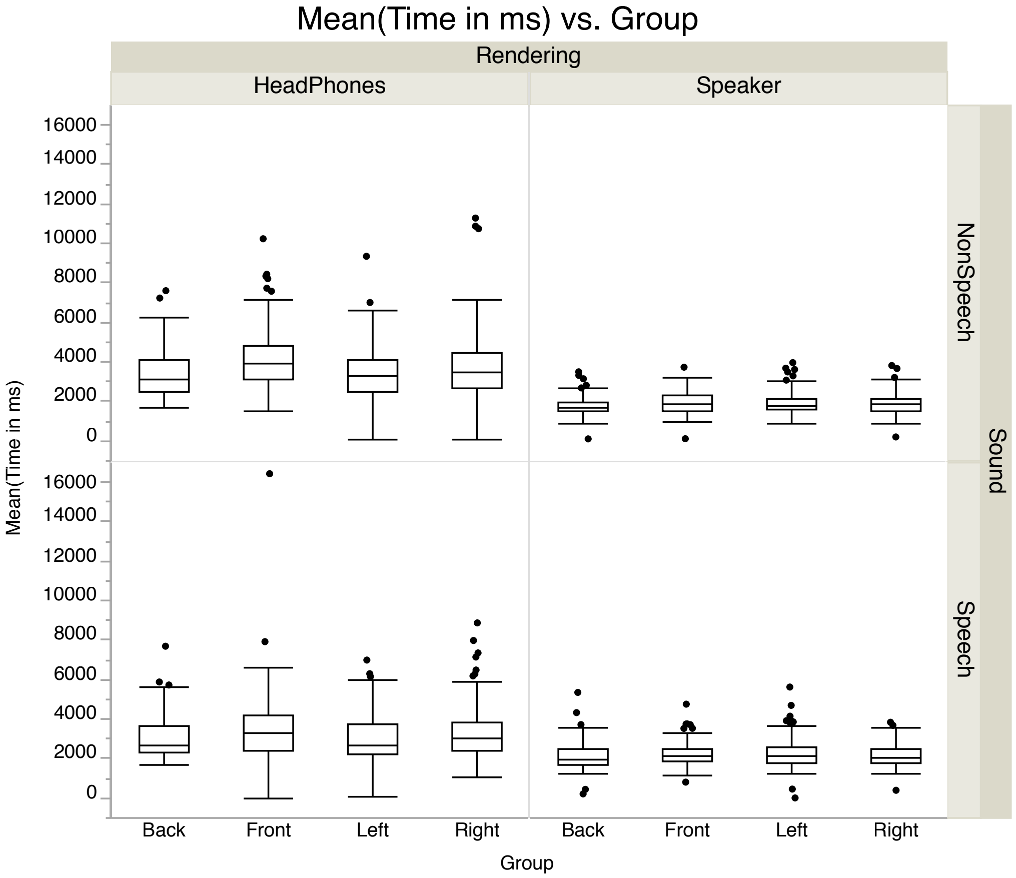


Figure 4.5: Mean time of orientating against the four grouped directions.

not show any significant differences between *left* and *right* orientation. This means that there is no difference in time of orientating to the left or to the right, or vice-versa. Users were not confused from sources of another direction. Also, the results of [Bronkhorst, 1995] showed that left-right confusion never occurred, which is also true in our case.

No left-right confusion discovered.

4.4 Second Experiment: Orientation Measurement and Presence

In the first experiment we analyzed the data we got from the Vicon tracking system. In this experiment we wanted to investigate how users perceive the virtual audio space when the orientation is tracked at different positions. Amongst other, real-world implementations use different (lower resolution) tracking systems, too. Therefore we let users fill out a questionnaire.

4.4.1 Conditions & Methodology

Placement and description of used audio files.

For this experiment, we placed six visual sound sources in the Coronation Hall (Figure 3.4). The sound sources spanned an area of 8 × 28 meters. The Coronation Hall has a size of 20 × 45 meters. We did not used the margin of the room, because the tracking at this area was not quite accurate.

Participants were holding a device on which the sound source they had to go to was displayed. Each sound source represented one of the following subjects: colors, fruits, drinks, names, cities, animals. We used a text-to-speech tool to create audio clips of the subjects. After entering the 2 m capture radius the user was notified by a sound-sample and a short red flash on the screen that they had reached the designated source. Every user had to reach all six sound sources in a given sequence to complete a task.

Explanation of the users tasks.

We created three different paths, one for every compass placement and randomized the order for every user. After every trial they had to fill out a questionnaire (Figures A.1 - A.4). The questions were taken from the questionnaire of [Witmer and Singer, 1998]. The compass was then mounted at a different position for the next trial (head / body / device).

We used the same compass for all positions. The order of placement of the compass was also randomized. So users would walk different paths while having the same compass placement to antagonize influences of the paths depending on the compass placement. Before the experiment started, each participant had to walk through the audio space to get an impression of the system.

4.4.2 Results

9 users, 2 female, in the age of 20 - 25 (average 24) participated in this study. All questions were answered on a 7 point Likert scale, with 1 being the lowest and 7 the highest score. We will only present some of the questions because the whole questionnaire has too many (Figure 4.6).

No significant difference was found between different tracking types.

We did not find any significant difference between the different tracking types. Nonetheless, head tracking gets the overall best score ($M = 5.15, SD = 1.6$) followed by device ($M = 4.88, SD = 1.45$) and body tracking ($M = 4.78, SD = 1.58$). The question *How natural did your interactions with the environment seem?* head- ($M = 5.875, SD = 0.83$), device- ($M = 5.44, SD = 1.13$), and body tracking ($M = 5, SD = 1.32$) were all rated in average better or equal than 5. For the responsiveness question head tracking ($M = 5.75, SD = 0.89$) received the best results before device- ($M = 5.33, SD = 1.41$) and body tracking ($M = 5.22, SD = 1.2$). The question *How well could you localize sounds?* was rated with ($M = 5.875, SD = 0.99$) for head, ($M = 5.11, SD = 1.36$) for device, and ($M = 4.44, SD = 1.51$) for body. The perceived stability of sound sources was rated slightly higher for head tracking ($M = 6.13, SD = 0.99$) than for device ($M = 5.33, SD = 1.23$) and body ($M = 4.44, SD = 1.88$).

4.4.3 Discussion

We assume that device tracking could be an alternative.

In all conditions head tracking was rated better than device tracking. The difference between the means of head- and device tracking is less then 1 point in every case. Since there is no significant effect between the different renderings, we assume that device tracking could be an alternative in such virtual sound spaces. Just keep in mind that the precision in this study was not as relevant as in the first study. The capture radius, that indicates if you reached a sound source was set to 2 meters. So if the precision is not that important, we assume that device tracking would be precise enough to not effect the users experience.

The mean task completion time of the different compass

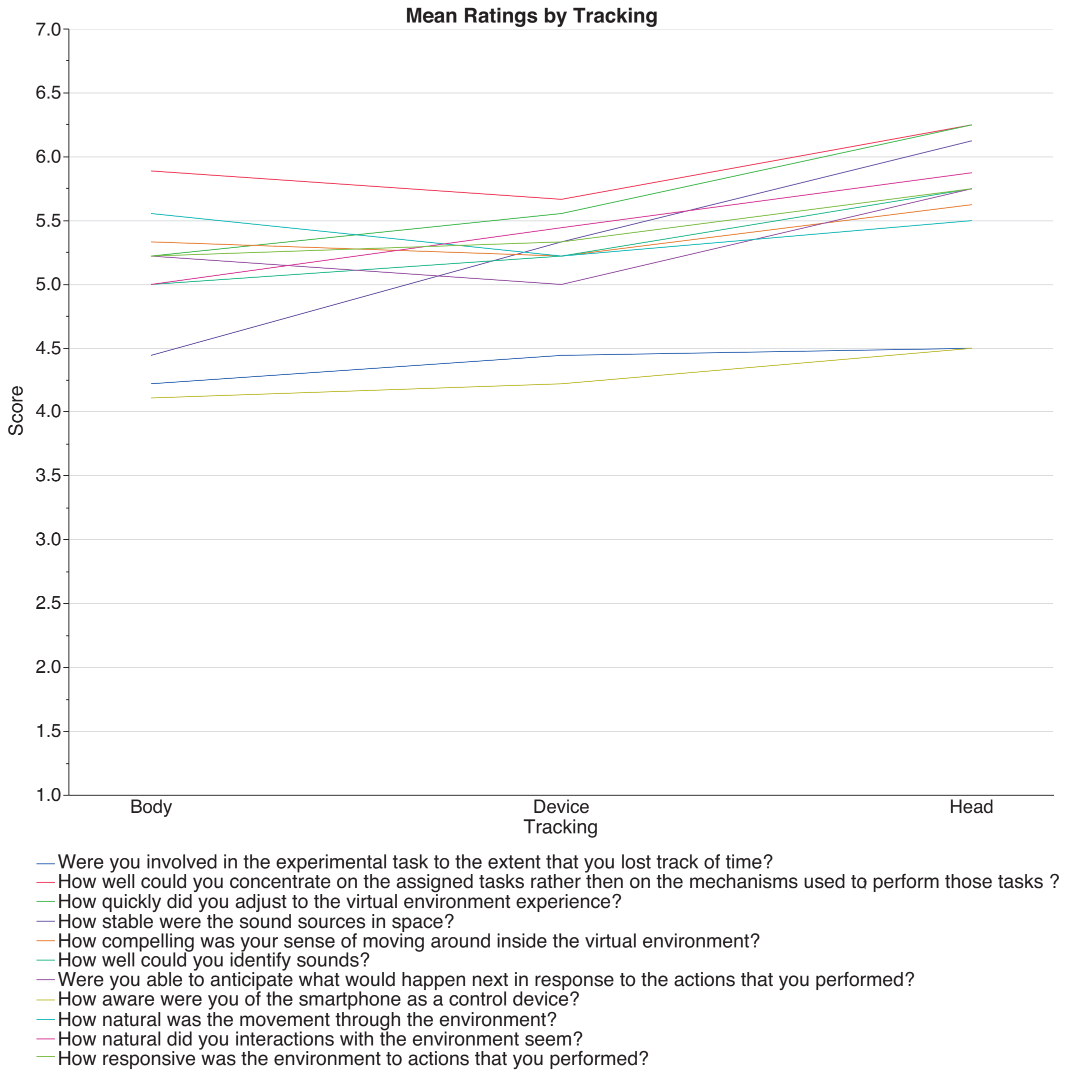


Figure 4.6: Most interesting questions of our perceived presence questionnaire by tracking.

placements shows a similar result. Head tracking ($M = 192s$) was fastest, followed by device tracking ($M = 198s$) and than with a bigger difference body tracking ($M = 245s$).

Chapter 5

Summary and Future Work

In this thesis we investigated how users orient towards sound sources and whether the orientation tracking could be done at a different place than the head. Many of the currently existing installations are using the head for orientation tracking. This needs additional hardware to be mounted onto the users head. Therefore we conducted two studies. The first study was performed under lab-settings. Participants had to orient towards sound sources with two different renderings and sound-types. The second study was conducted in a real world setting using *Corona*. We let participants do walking tasks with three different tracking sensor placements. Afterwards, participants had to fill out a perceived presence questionnaire.

5.1 Summary and Contributions

The first study shows that users do initial head-turns at the beginning to orient in the first 600 ms. These head turns are needed to have an initial orientation. If one removes these initial head turns, e.g., by using device tracking, the perceived latency will be high enough to be noticeable but completing navigational tasks will still be possible Mariette

[2010].
We also found the phenomenon of front-back confusion, which is a common problem in virtual sound spaces. Users take longer to orient towards sound sources in front of them than behind them.

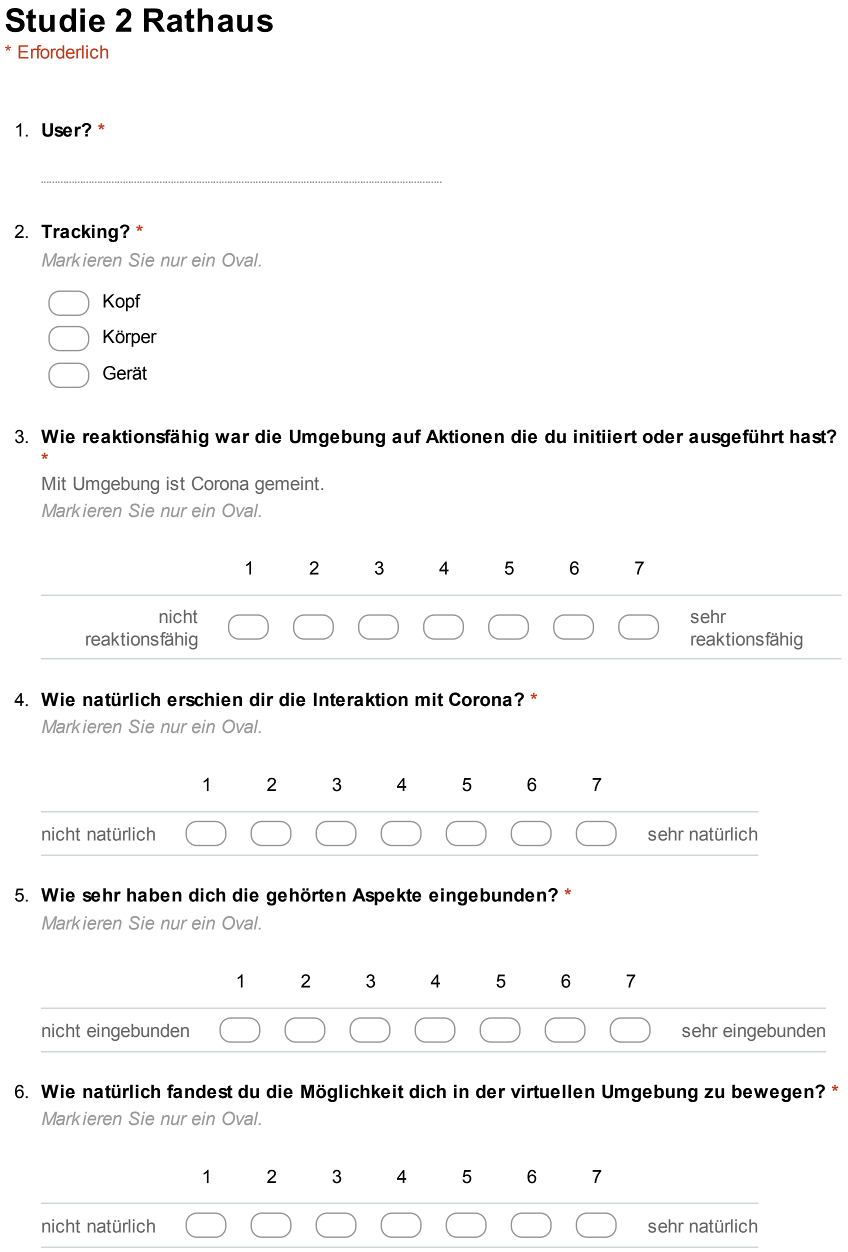
The results of the second experiment show that there is no major difference for the user if we use other tracking positions than head, assuming that precision is not the key.
Based on all these results we suggest to use head tracking if precision is important. Otherwise device-tracking would be a good alternative.

5.2 Future Work

In the second study we did not tell users which position is used for tracking. An interesting study would be to tell participants which position is tracked and then analyze their behavior if there is any difference to our actual data.
Although, one could test the hardware of current smart phones if the desired accuracy and delay mentioned by Mariette [2010] and [Walker and Lindsay, 2006] is given.

Appendix A

Presence Questionnaire



Studie 2 Rathaus

* Erforderlich

1. **User?** *

   ____

2. **Tracking?** *

   Markieren Sie nur ein Oval.

   ○ Kopf
   ○ Körper
   ○ Gerät

3. **Wie reaktionsfähig war die Umgebung auf Aktionen die du initiiert oder ausgeführt hast?** *

   Mit Umgebung ist Corona gemeint.

   Markieren Sie nur ein Oval.

	1	2	3	4	5	6	7	
nicht reaktionsfähig	○	○	○	○	○	○	○	sehr reaktionsfähig

4. **Wie natürlich erschien dir die Interaktion mit Corona?** *

   Markieren Sie nur ein Oval.

	1	2	3	4	5	6	7	
nicht natürlich	○	○	○	○	○	○	○	sehr natürlich

5. **Wie sehr haben dich die gehörten Aspekte eingebunden?** *

   Markieren Sie nur ein Oval.

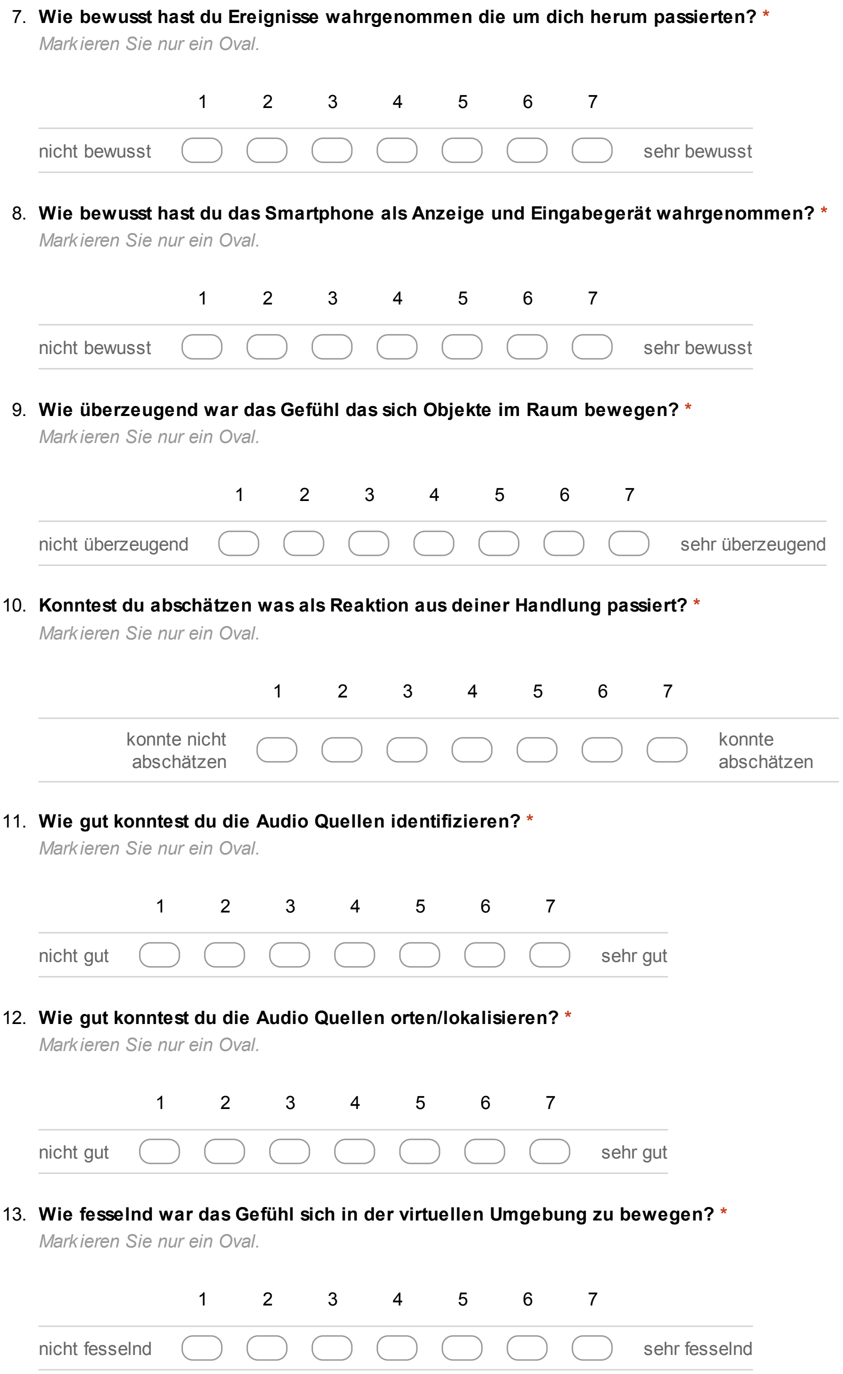
	1	2	3	4	5	6	7	
nicht eingebunden	○	○	○	○	○	○	○	sehr eingebunden

6. **Wie natürlich fandest du die Möglichkeit dich in der virtuellen Umgebung zu bewegen?** *

   Markieren Sie nur ein Oval.

	1	2	3	4	5	6	7	
nicht natürlich	○	○	○	○	○	○	○	sehr natürlich

Figure A.1: The perceived presence questionnaire the user had to fill out after every task during the second study (page 1).



7. **Wie bewusst hast du Ereignisse wahrgenommen die um dich herum passierten?** *

Markieren Sie nur ein Oval.

	1	2	3	4	5	6	7	
nicht bewusst	○	○	○	○	○	○	○	sehr bewusst

8. **Wie bewusst hast du das Smartphone als Anzeige und Eingabegerät wahrgenommen?** *

Markieren Sie nur ein Oval.

	1	2	3	4	5	6	7	
nicht bewusst	○	○	○	○	○	○	○	sehr bewusst

9. **Wie überzeugend war das Gefühl das sich Objekte im Raum bewegen?** *

Markieren Sie nur ein Oval.

	1	2	3	4	5	6	7	
nicht überzeugend	○	○	○	○	○	○	○	sehr überzeugend

10. **Konntest du abschätzen was als Reaktion aus deiner Handlung passiert?** *

Markieren Sie nur ein Oval.

	1	2	3	4	5	6	7	
konnte nicht abschätzen	○	○	○	○	○	○	○	konnte abschätzen

11. **Wie gut konntest du die Audio Quellen identifizieren?** *

Markieren Sie nur ein Oval.

	1	2	3	4	5	6	7	
nicht gut	○	○	○	○	○	○	○	sehr gut

12. **Wie gut konntest du die Audio Quellen orten/lokalisieren?** *

Markieren Sie nur ein Oval.

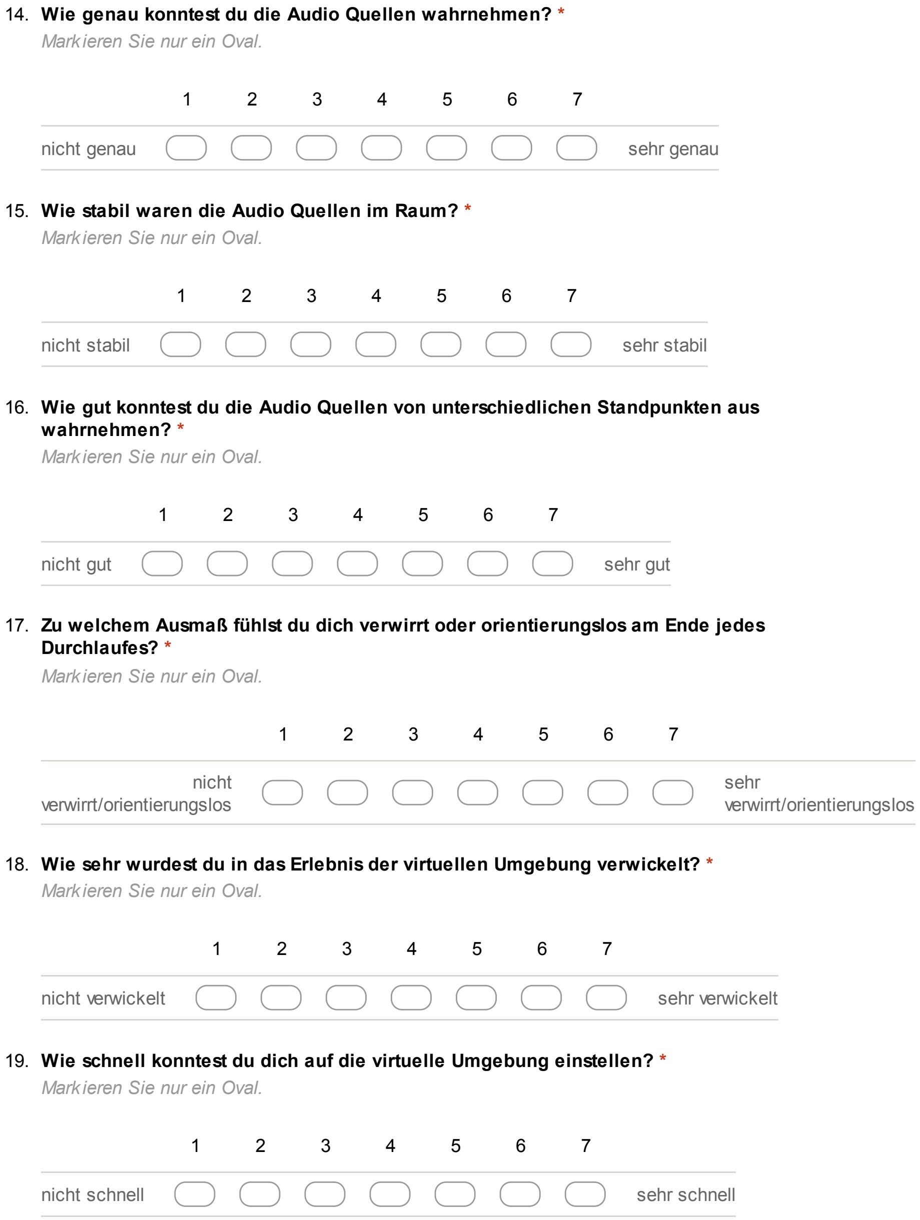
	1	2	3	4	5	6	7	
nicht gut	○	○	○	○	○	○	○	sehr gut

13. **Wie fesselnd war das Gefühl sich in der virtuellen Umgebung zu bewegen?** *

Markieren Sie nur ein Oval.

	1	2	3	4	5	6	7	
nicht fesselnd	○	○	○	○	○	○	○	sehr fesselnd

Figure A.2: The perceived presence questionnaire the user had to fill out after every task during the second study (page 2).



14. **Wie genau konntest du die Audio Quellen wahrnehmen?** *
Markieren Sie nur ein Oval.

	1	2	3	4	5	6	7	
nicht genau	○	○	○	○	○	○	○	sehr genau

15. **Wie stabil waren die Audio Quellen im Raum?** *
Markieren Sie nur ein Oval.

	1	2	3	4	5	6	7	
nicht stabil	○	○	○	○	○	○	○	sehr stabil

16. **Wie gut konntest du die Audio Quellen von unterschiedlichen Standpunkten aus wahrnehmen?** *
Markieren Sie nur ein Oval.

	1	2	3	4	5	6	7	
nicht gut	○	○	○	○	○	○	○	sehr gut

17. **Zu welchem Ausmaß fühlst du dich verwirrt oder orientierungslos am Ende jedes Durchlaufes?** *
Markieren Sie nur ein Oval.

	1	2	3	4	5	6	7	
nicht verwirrt/orientierungslos	○	○	○	○	○	○	○	sehr verwirrt/orientierungslos

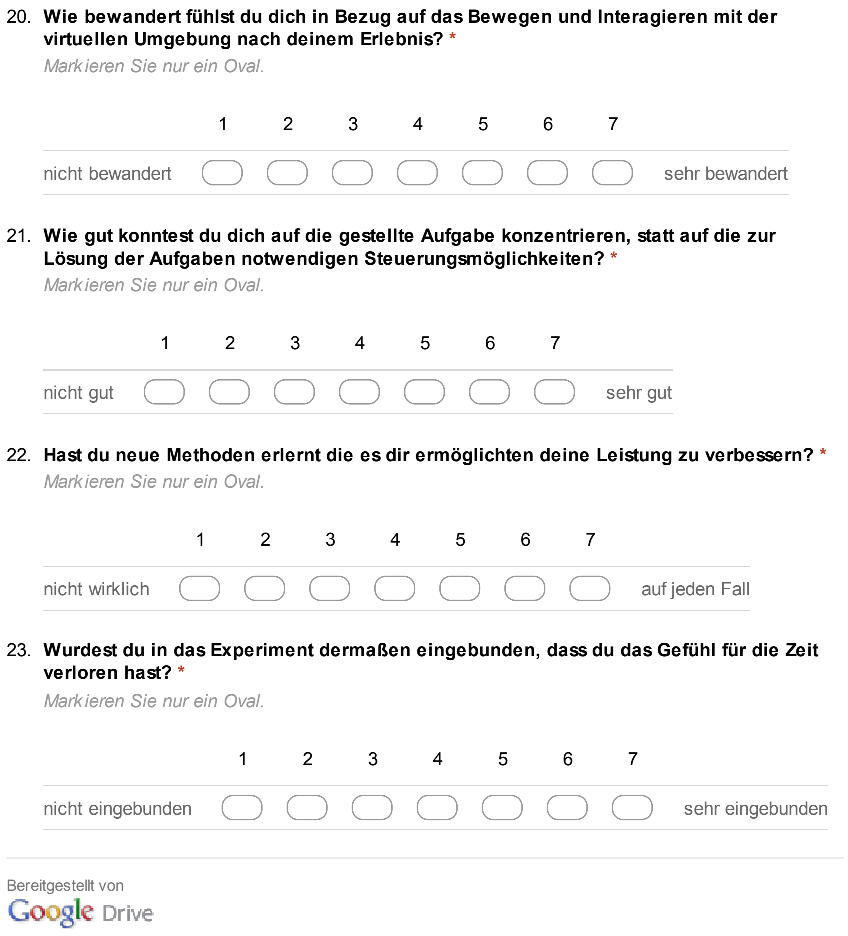
18. **Wie sehr wurdest du in das Erlebnis der virtuellen Umgebung verwickelt?** *
Markieren Sie nur ein Oval.

	1	2	3	4	5	6	7	
nicht verwickelt	○	○	○	○	○	○	○	sehr verwickelt

19. **Wie schnell konntest du dich auf die virtuelle Umgebung einstellen?** *
Markieren Sie nur ein Oval.

	1	2	3	4	5	6	7	
nicht schnell	○	○	○	○	○	○	○	sehr schnell

Figure A.3: The perceived presence questionnaire the user had to fill out after every task during the second study (page 3).



20. **Wie bewandert fühlst du dich in Bezug auf das Bewegen und Interagieren mit der virtuellen Umgebung nach deinem Erlebnis?** *
Markieren Sie nur ein Oval.

	1	2	3	4	5	6	7	
nicht bewandert	○	○	○	○	○	○	○	sehr bewandert

21. **Wie gut konntest du dich auf die gestellte Aufgabe konzentrieren, statt auf die zur Lösung der Aufgaben notwendigen Steuerungsmöglichkeiten?** *
Markieren Sie nur ein Oval.

	1	2	3	4	5	6	7	
nicht gut	○	○	○	○	○	○	○	sehr gut

22. **Hast du neue Methoden erlernt die es dir ermöglichten deine Leistung zu verbessern?** *
Markieren Sie nur ein Oval.

	1	2	3	4	5	6	7	
nicht wirklich	○	○	○	○	○	○	○	auf jeden Fall

23. **Wurdest du in das Experiment dermaßen eingebunden, dass du das Gefühl für die Zeit verloren hast?** *
Markieren Sie nur ein Oval.

	1	2	3	4	5	6	7	
nicht eingebunden	○	○	○	○	○	○	○	sehr eingebunden

Bereitgestellt von
Google Drive

Figure A.4: The perceived presence questionnaire the user had to fill out after every task during the second study (page 4).

Bibliography

Anupriya Ankolekar, Thomas Sandholm, and Louis Yu. Play it by ear: a case for serendipitous discovery of places with musicons. In *Proc. CHI '13*, pages 2959–2968. ACM, 2013. ISBN 978-1-4503-1899-0. doi: 10.1145/2470654.2481411.

Adelbert W. Bronkhorst. Localization of real and virtual sound sources. *The Journal of the Acoustical Society of America*, 98(5):2542–2553, 1995. doi: 10.1121/1.413219.

Florian Heller, Thomas Knott, Malte Weiss, and Jan Borchers. Multi-user interaction in virtual audio spaces. In *Proc. CHI EA '09*. ACM, 2009. doi: 10.1145/1520340.1520688.

Florian Heller, Aaron Krämer, and Jan Borchers. Simplifying orientation measurement for mobile audio augmented reality applications. In *(to appear) Proc. CHI '14*. ACM, Apr 2014. doi: 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), January 2002.

Georgios N Marentakis and Stephen A Brewster. Effects of feedback, mobility and index of difficulty on deictic spatial audio target acquisition in the horizontal plane. In *Proc. CHI '06*, pages 359–368. ACM, 2006. ISBN 1-59593-372-7. doi: 10.1145/1124772.1124826.

Nicholas Mariette. Navigation performance effects of render method and head-turn latency in mobile audio augmented reality. In *Proc. ICAD '09*, pages 239–265.

Springer, Copenhagen, 2010. ISBN 978-3-642-12438-9. doi: 10.1007/978-3-642-12439-6_13.

Christoph Stahl. The roaring navigator: a group guide for the zoo with shared auditory landmark display. In *Proc. MobileHCI '07*. ACM, 2007. doi: 10.1145/1377999.1378042.

Tuyen V Tran, Tomasz Letowski, and Kim S Abouchacra. Evaluation of acoustic beacon characteristics for navigation tasks. *Ergonomics*, 43(6):807–827, 2000. doi: 10.1080/001401300404760.

Yolanda Vazquez-Alvarez, Ian Oakley, and StephenA Brewster. Auditory display design for exploration in mobile audio-augmented reality. *Personal and Ubiquitous computing*, 16(8):987–999, 2012. doi: 10.1007/s00779-011-0459-0.

Bruce N Walker and Jeffrey Lindsay. Navigation Performance With a Virtual Auditory Display: Effects of Beacon Sound, Capture Radius, and Practice. *Human Factors*, 48 (2):265–278, 2006. doi: 10.1518/001872006777724507.

Bob G Witmer and Michael J Singer. Measuring Presence in Virtual Environments: A Presence Questionnaire. *Presence: Teleoper. Virtual Environ.*, 7(3):225–240, 1998. doi: 10.1162/105474698565686.

Index

Typeset April 2, 2014