RWTH AACHEN UNIVERSITY

CORONA

Implementation and Evaluation of Continuous Virtual Audio Spaces for Interactive Exhibits

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

by
Thomas Knott

Thesis advisor:
Prof. Dr. Jan Borchers

Second examiner:
Prof. Dr. Torsten Kuhlen

Registration date:	Sep 30th, 2008
Submission date:	Jun 15th, 2009

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, Jun15th, 2009
Thomas Knott

Contents

List of Figures

List of Tables

Abstract

Many historic sites are rich of history and have witnessed important events of former times. Some of these events did not leave visual traces of their existence. To convey an impression of these events, they have to be made noticeable for the visitor of the site. CORONA, a project part of the Route Charlemagne, is an implementation of such a system in the Coronation Hall of the city hall of Aachen, Germany. It is planned to be an interactive multimedia experience, which conveys information to the visitor in an innovative and compelling way. The idea of CORONA is to let the visitor immerse in a virtual audio space and by that to let the visitor witness a historical event.

This thesis will give an overview of related touristic projects in the field of interactive audio spaces. After that we describe the development of an interactive prototype which integrates a position tracking and a spatial audio rendering. It enables the creation and the test of virtual audio spaces. Furthermore, this work describes the implementation and the evaluation of several audio rendering methods, which are adapted to the technical systems of the target application. In this context we explain the relevant fundamentals of sound propagation and spatial hearing, which are essential to understand this work. Another part is dedicated to the different ideas of adapting the audio rendering to the specific requirements, which are induced by the application of the system as an interactive exhibit.

Überblick

Viele historische Stätten sind reich an Geschichte und wurden Zeuge wichtiger Ereignisse vergangener Zeiten, welche keine visuellen Spuren hinterlassen haben. Um einen Eindruck dieser Ereignisse vermitteln zu können, müssen diese für den Besucher wahrnehmbar gemacht werden. Dieses soll für den Krönungssaal des Aachener Rathauses mithilfe von CORONA, einem Teilprojekt der Route Charlemagne, umgesetzt werden. CORONA ist geplant als ein interaktives Multimedia-Erlebnis, welches Informationen in einer innovativen und für den Besucher spannenden Form vermittelt. Die Idee von CORONA ist, die Besucher in eine virtuelle Klangwelt eintauchen zu lassen, welche es ihnen ermöglicht ein längst vergangenes historisches Ereignis mitzuerleben.

Diese Arbeit gibt zuerst einen Überblick über verwandte touristische Projekte auf dem Gebiet der interaktiven Klangwelten. Daraufhin beschreiben wir die Entwicklung eines interaktiven Prototypen, welcher es ermöglicht virtuelle Klangwelten zu erstellen und in diese, unter Einbindung von Systemen zur Positionsbestimmung und räumlicher Klangsynthese, einzutauchen. Weiterhin beschreibt diese Arbeit die Entwicklung und die Tests von mehreren, an die im Projekt verwendeten technischen Systeme angepassten, Klangsynthese-Methoden. In diesem Zusammenhang werden die zum Verständnis der Arbeit relevanten Grundlagen der Schallausbreitung und des räumlichen Hörens beschrieben. Ein weiterer Bereich der Arbeit widmet sich verschiedenen Ideen zur Anpassung der Klangsynthese an die spezifischen Anforderungen, welche durch die Verwendung des Systems als interaktives Exponat entstehen.

Conventions

Throughout this thesis we use the following conventions.

Unidentified third persons are always described in female form for the purpose of political correctness.

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*

The whole thesis is written in American English.

Chapter 1

Introduction

The idea behind CORONA is to create an interactive experience in the Coronation Hall situated in the historic city hall of Aachen, Germany. CORONA is part of the Route Charlemagne, which is a project to make Aachen more interesting for strangers and visitors. The Route has several stations and leading through the whole city and treat several important topics like "History", "Science", "Religion", "Media", "Europe", "Economy", and, "Power". Exhibitions in historical sites, architectural works, innovational media and cultural events should bring these topics to a broad public. The Route is still in a developing process and therefore it is still expanded by innovational ideas and new topics. One of the first stations that will be affiliated to the Route is the city hall of Aachen and its Coronation Hall representing the topic "Power".

CORONA is part of the Route Charlemagne

The city hall was built in the early 14th century by the residents of Aachen as a sign for their freedom as citizens. As a concession to the crown, they obliged to built the Coronation Hall as a place where the feast after the coronation ceremony in the Aachen Cathedral could be celebrated. Until today, it has not lost its importance as venue for var-

Coronation Hall of the city hall of Aachen, Germany ious cultural events. As the most famous one the International Charlemagne Prize of the city of Aachen may be mentioned.

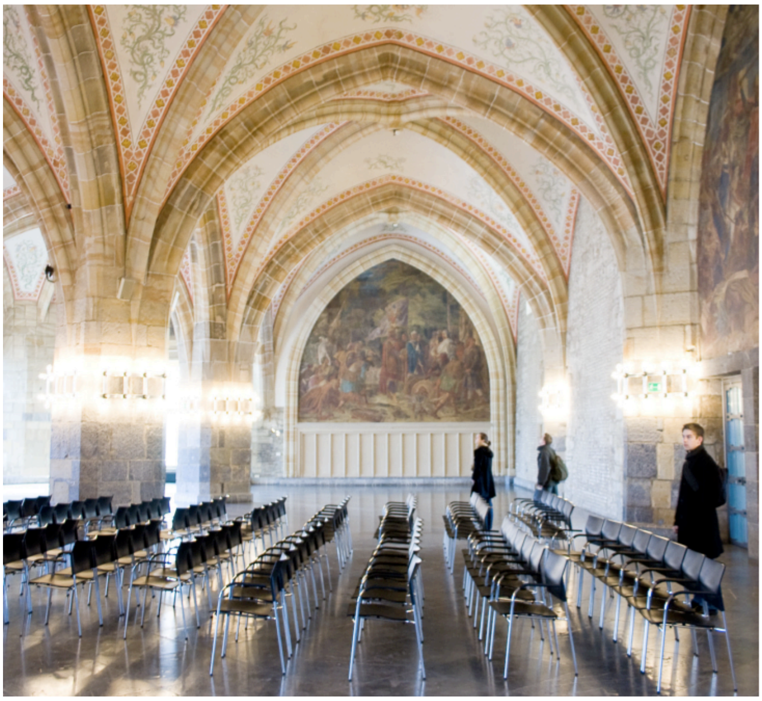


Figure 1.1: The Coronation Hall.

Reveal hidden history

The Coronation Hall is rich of history, which is not visible to the visitor. Since the coronation hall is still frequently in use, it is not possible to set up permanent installations in the hall. To uncover hidden history and to convey it to a broad public, the idea came up to create a completely virtual audio space which augments the coronation hall with an additional layer of information. This interactive experience is planned to revive a medieval coronation feast.

Witnessing the coronation ceremony

In this virtual scene a set of historic characters is included and these personalities provide information by conversing with each other. For example King Charles V discusses the situation of the people in the time of the plague with the archbishop of Trier. These conversations convey historic information in a vivid and compelling way. By witnessing the coronation ceremony, the visitor gets an insight into the conventions of former times.

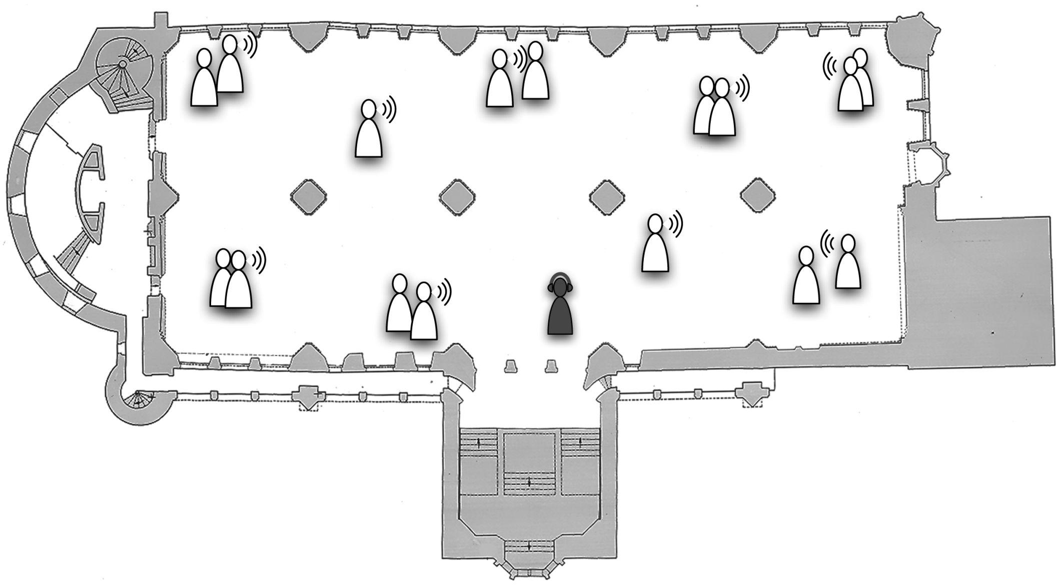


Figure 1.2: Floor plan of the Coronation Hall with virtual speakers (white) and a visitor (black).

Acoustic simulation of a coronation feast

The audio space is an acoustic simulation of a coronation feast containing several virtual sound sources, corresponding to the historic characters. The sound sources are distributed in the whole coronation hall and rendered on a mobile device which is handed out to the visitors. The rendering process incorporates the position and orientation of the listener as well as the position of the virtual sound sources to emulate a natural sound propagation. The visitors experience this continuous virtual audio space only by moving in the coronation hall. As there are no visual representations of the sound sources, the visitors explore the virtual scene solely by listening and following the auditive cues of the sound sources [Heller et al. [2009]].

Explore virtual scenes solely by listening

Overview

This thesis is about the design, implementation and evaluation of continuous virtual audio spaces in the context of CORONA. In the following we will give a brief overview on the structure of this thesis.

Related work

Chapter 2 gives an overview of existing audio guide technologies used to augment exhibition spaces with additional information. Furthermore we will describe several research projects in the area of virtual audio experiences and audio augmented reality.

First interactive prototype

Chapter 3 describes the implementation and evaluation of a first interactive prototype. We will describe the informations we wanted to gain with it and the resulting requirements which have to be fulfilled by the prototype. After that we will give a brief insight into some implementation details and a description of an informal evaluation. Finally we will describe why we came to the conclusion that we have to further improve our spatial audio rendering.

Foundations of auralization

Therefore, Chapter 4 will explain the basics of sound propagation and spatial hearing. In this chapter we will describe the foundations of spatial audio rendering which are needed to understand the approaches described in the next chapter.

Implementation

Chapter 5 explains the implementation of three spatial audio rendering methods and how we incorporated a simple room acoustic into our audio space. Afterwards, we will report on a preliminary evaluation and the obtained results. Finally, we will describe some further changes and ideas which were stimulated by the study.

Evaluation

The final evaluation of our implemented audio rendering methods is described in Chapter 6. Additionally, we will test two ideas to improve the listening experience when being near a virtual sound source in this chapter.

Chapter 7 summarizes the goals and contributions of this thesis and will give an overview of future work which has to be done.

Chapter 2

Related Work

2.1 Audio Guides

The idea of augmenting exhibition spaces and historic sites with additional audio information by so called audio guides appeared almost half a century ago. It began in 1957 with the Acoustiguide, guiding through the home of President D. Roosevelt by a narration of his widow using a carried along 35 cm long reel to reel cassette player [Maryanne Leigh [2007]]. In the first years, analog playback devices were utilized. They allowed, beside of controlling the volume, pausing or rewinding the tapes, no further interaction with the content. Constrained by the linear access to the information, listeners had to follow a particular path through the exhibition. In those days, most places offered only one taped tour which was technically limited to 45 minutes and at least had to fit the interests and the preferred detail level of all visitors.

Linear access to information

This problems were solved in the late 90's by adopting the general technological progress and by incorporating new random access audio storage. A common variant were the keypad-based audio guides. While walking through an exhibition on a self-chosen path, the user could gain information about the actually viewed exhibits by entering a number, which was attached to the item. Besides the advantage of choosing an own chronology and walking with a self-

Keypad-based audio guides

determined pace, the visitor had the additional possibility to select the conveyed information according to her interests and thereby personalize the experience. This simple but effective solution is still used nearly at every big museum up to nowadays [Proctor and Tellis [2003]].
Nevertheless this variant has drawbacks, e.g., if a visitor wants to gain information about a specific exhibit, she first has to find the number according to the item in the exhibition space. This could especially become a problem for large objects or objects which could not directly be labeled like e.g. paintings on a ceiling. If the code has been found it furthermore has to be read and entered into the audio guide. This keypad-based interaction may pull the visitor's attention away from the actual exhibits and may disturb the overall experience flow.

Drawback: pull attention away from exhibits

An audio guide type avoiding these problems was developed as wireless headphone technology evolved in the eighties. Induction or infrared-based techniques were used to broadcast exhibit descriptions continuously to the local area around the exhibit. When the visitor came close to the exhibit, she entered the range of the corresponding sender and received the audio content on her headphones. By this, the listener could access information by solely walking through the exhibition space without the need for an additional input device or interaction [Eckel [2001]].
As there was only one audio broadcast per exhibit, all visitors at one exhibit synchronously receive the same looping signal and share the same timing of the broadcasted audio sequences. The obvious drawback of this system was, that a visitor may arrive in the middle of an already started audio description.
Nevertheless "the reason many people go to museums is to socialize, to be with friends and to discuss the exhibit as they experience it" [Bederson [1995]]. Thereby sharing the experience with the companions is often a higher priority than education [Hood [1983]]. These social goals were supported since companions only had to stay in proximity to each other to share the same audio experience.

Infrared based technology

Access information solely by walking

Socialize with companions

A tour guide which tried to combine the advantages of the latter two audio guide variants is described in Bederson [1995]. Like in the keypad-based variant, the visitor had to carry a random access audio device and the audio comments of the exhibits were associated with a unique code. But instead of being written on labels in the exhibition space, the codes were locally broadcasted by small infrared transmitters installed above every exhibit into its direct neighborhood. Furthermore an audio guide was equipped with a microprocessor and an infrared receiver. Coming close to an exhibit, the guide automatically started the associated comment and stopped if the visitor walked away.

Automatic start when reaching an exhibit

Exhibit information can again be explored by solely walking through the exhibition space but this time without the disadvantage of hopping into an already started clip. Since the starting and stopping of the audio clips only was depending on the visitors' position, companions approaching an exhibit side by side would share an identical audio experience.

The audio clips of all the so far described systems had a binary playing state – the clip was either audible or not. The designers of the following system took a different approach.

2.2 ec(h)o

The ec(h)o system is an "augmented reality interface" which overlays the exhibition space of the Canadian Museum of Nature in Ottawa with a three- dimensional soundscape [Ron et al. [2004]]. Headphones and a combination of RFID-based and optical tracking of the visitor's position are used to present a dynamic audio experience to the visitor. The soundscape creates an atmosphere corresponding to the artifacts next to the visitor. It consists of an abstract ambient part and a part with short audio sequences. The abstracts ambience relates to the overarching theme of the exhibits currently surrounding the visitor, e.g., when being among marine exhibits the sound of the sea could be heard over the headphones. Moving through the exhibition space and between different themes the ambiences fade in and out continuously according to the related artifacts in the visitor's proximity (see figure 2.1 a). Additionally, short audio sequences associated with a specific artifact in the visi-

Three-dimensional soundscape

Ambiences fade in and out continuously

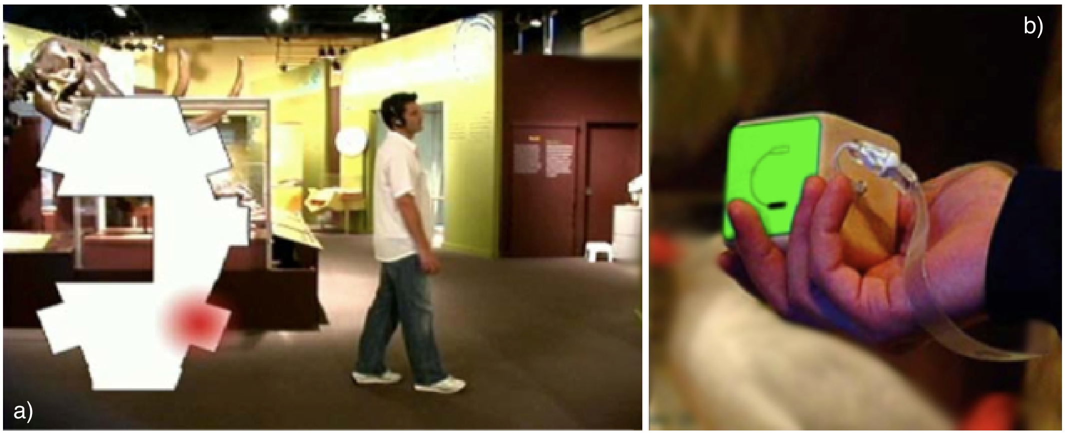


Figure 2.1: a) A visitor walking through an exhibit room. On the left hand one can see a room map, the red gradient depicting the volume of an ambient sound, heard when being on the map position. b) The tangible user interface. (both figures taken from Wakkary and Hatala [2007])

tors optical range are played once in a while to engage her to take a closer look, e.g., the cry of a sea gull should draw attention to a sea gull preparation. Because the visitor's orientation is not tracked by the system there is no spatial audio synthesis.

Tangible input device

A second mode of interaction let the visitor retrieve more information about nearby artifacts by communicating with the system through gestures performed with a carried along tangible object,which has the form of a cube (see 2.1 b). Coming close to an item, three related acoustic prefaces are played to the visitor. The first is thereby solely presented on the left ear channel, the second on left and right, and the third solely on the right ear channel. By turning the cube into one of the directions which are hinted by the sound panning left and right, the according preface is selected and an associated answer is played. When the answer of the audio sequence ends and the visitor is still near the artifact, the system presents a new set of preludes and a kind of conversation emerges. Every time new prefaces are offered they are dynamically selected by an agent system. The selection process is based upon past interactions and is influenced by the individual interests of the visitor. To contrast with the more formal answers and to create a sense of surprise, discovery, and play, the prefaces use varied types of riddles, word plays and puns, e.g., the pref-

Influence of individual interests

ace: "Longer than you would want to know" and following answer: "Tapeworms come in varying lengths and sizes. Interestingly, the longest recorded tapeworms have been those that live in humans" [Wakkary and Hatala [2007]].

Additional auditive cues

As already mentioned, the ambience part of the soundscape provides the visitor with information about the general topic of the surrounding items. Additionally, since the loudness of the theme of the ambience clip is mapped to the relative visitor-to-artifact distance, the visitor may also infer, e.g., whether a specific theme is approached or left behind. The additionally given hints about nearby special artifacts transports information about their local presence but do not facilitate the localization of the object in the exhibition space. The following system tried to close this gap.

2.3 The Roaring Navigator

Lightweight navigational aid

Reduce visitors' isolation

The aim of the Roaring Navigator project is to create an electronic tour guide offering two main functionalities in a pure audible manner: (first) the augmentation of the exhibits with additional details and (second) an auditory landmarks display, providing spatial survey knowledge and "lightweight" navigational aid [Stahl [2007]]. Its secondary goal is to minimize the isolation effect between companions, caused by the presentation of audio content through headphones. The project is situated in a zoo environment and is implemented on PDAs from HP, which are combined with Bluetooth GPS receivers and digital compasses. For the group mode the PDAs are connected pairwise by wifi ad-hoc networks.

Auditory landmark display

The auditory landmark display informs the visitor about points of interest, the so called landmarks, in the visitors near environment. Therefore, every landmark is associated with a characteristic audio sequence, reflecting the object available on that point, e.g., an animal enclosure or a restaurant. While the visitor moves through the zoo, the system continuously presents her the audio sequences of the three nearest landmarks in a random order (see figure

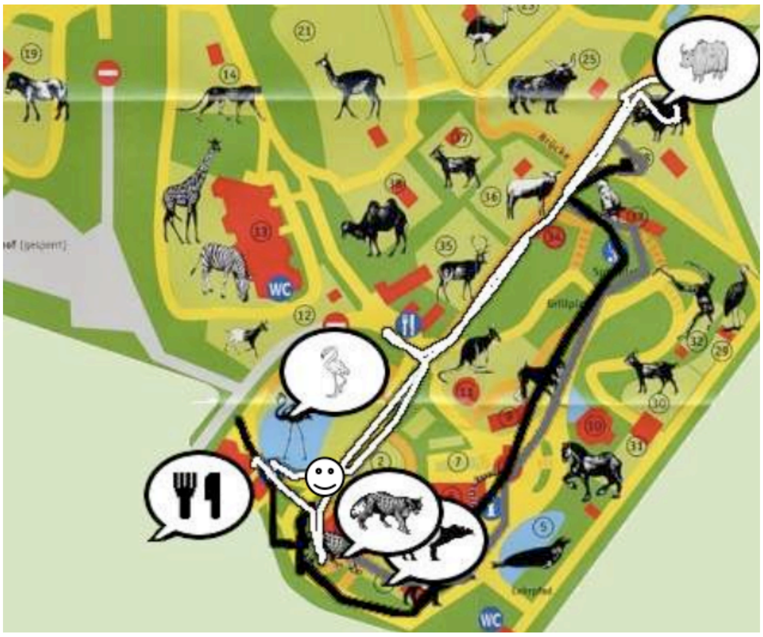


Figure 2.2: Map of the zoo with some auditory landmarks (taken from Stahl [2007]).

2.2). Since the auditory landmark display shall be unobtrusive and not pull away the user's attention from the environment, recordings of animal voices are used so that the perceived soundscape naturally fits into the existing normal zoo soundscape. To give additionally hints about how to reach a point of interest, the sounds were stereo balanced and adjusted in volume, thus giving information about relative direction and distance to the visitor.

Stereo balanced sound

If the visitor moves closer to a point of interest for the first time an explanatory comment, e.g., "you are close to the gibbons", is presented, briefly explaining which landmark currently may be approached. When actually reaching the landmark, e.g., an enclosure, the system automatically delivers a detailed description of the visible object, in this case the animal.

Explanatory comment

To minimize the problem of the isolation of companions by the headphone presentation, a group mode is introduced in which the audio guides of group members are coupled in a way that all members hear the same audio sequences at the

Group mode

same time. In the single mode the sequences currently presented to the visitor are selected and manipulated only on the basis of her own position. To allow a synchronization of the heard content in the group mode, the PDA of one group member is configured as master device. By its position, the master PDA determines the selection and the scheduling of the played audio clips for all companions. The remaining other devices only manipulate the defined clips by stereo panning and volume adaption according to their own position.

Master device

2.4 LISTEN

LISTEN is a research project started in January 2001 and funded by the European Commission. It was planned to provide users a personalized and situated audio information space, augmenting a physical environment. [Eckel [2001]]. New forms of multi-sensory contents were proposed to create multimodal and immersive experiences for a broad spectrum of applications, ranging from art installations to entertainment events.

Personalized audio space

One application has been installed in the August Macke art exhibition at the Kunstmuseum Bonn [Terrenghi and Zimmermann [2004]]. The users wear motion-tracked wireless headphones for presentation of a 3D spatial reproduction of a virtual auditory scene. A sophisticated auditory rendering process incorporating an advanced binaural audio synthesis was used to integrate the virtual scene seamlessly with the real environment. Speech, music and sound effects are dynamically arranged, offering information related to visual objects placed in the exhibition and creating a context-specific atmosphere. The soundscape is not only personalized with respect to the spatial presentation, it also tries to adapt the selection of the presented information to better meet user's interests, preferences, and previous knowledge. A user-modeling-component uses the visitor's movements, the spatial history of the visit, and the interests expressed explicitly by the visitor to infer which information could be interesting for the user while watching an exhibit.

Advanced binaural audio synthesis

User-modeling-component

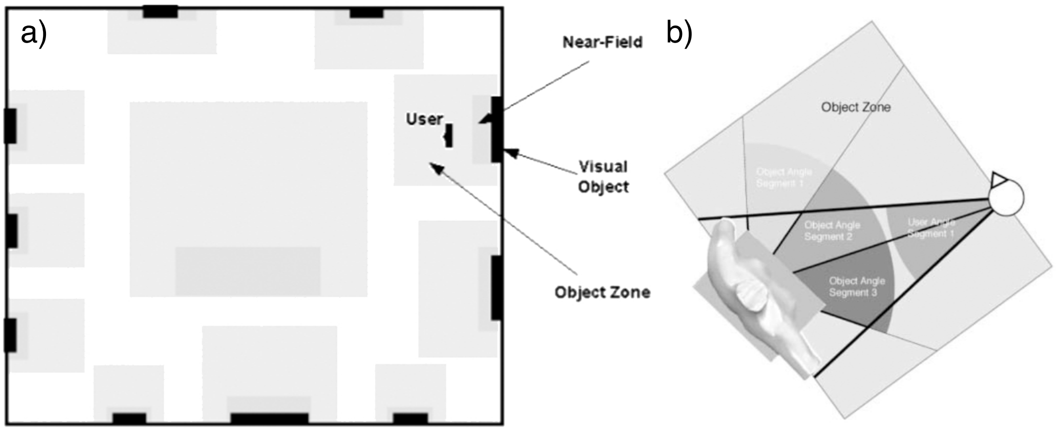


Figure 2.3: Example for the Objects Zones and Near Fields of the LISTEN Augmentation Layer (taken from Gossmann and Specht [2002] and Terrenghi and Zimmermann [2004]).

LISTEN uses four models

To implement this interaction, LISTEN uses four models: (1) a World Model, describing the physical exhibition space; (2) an Augmentation Layer on the World Model, defining areas in the World Model which contain active elements or sound objects with which the user interacts; (3) a Domain Model, describing the information about sound objects which are connected to physical objects by the Augmentation Layer; (4) a User Model, holding the visitor's profile and the visit history, deciding which of the sounds, associated via the Augmentation Layer with the actual position and orientation, are presented at one moment [Gossmann and Specht [2002]].

Augmentation Layer on World Model

Objects Zones and Near Fields

In figure 2.3 we can see two schematic examples depicting Augmentation Layers on World Models. In the figure 2.3 a) we see an exposition hall with several exhibits which is divided into several areas. Around every exhibit, a larger region defines the so called Object Zone. When entering this zone general information about the exhibit is presented to the visitor. By incorporating spatial audio synthesis methods the descriptions seems to be emitted directly by the exhibit. The Object Zones are further subdivided into Near Fields, which are connected to smaller parts of the physical object and contain more detailed sound information [Terrenghi and Zimmermann [2004]]. The selection of the presented information is not only depending on the listener's position but also on her relative orientation to the object. In figure 2.3 b) we can see an example of a statue's Near Fields. The statue has three different interesting perspectives with specific comment associated respectively. Therefore, the comments are only presented if the visitor is located in a specific angle segment and distance with respect to the statue. To present the information only if the visitor's locus of attention is potentially at the statue, the relative viewing direction has to be within a designated range, i.e., facing the object[Gossmann and Specht [2002]].

Perspectives with specific comments

The previously described systems have one thing in common, they all deliver information in a descriptive commentary style. A classic example for choosing a different way is the tour guide of Alcatraz Prison in San Francisco, USA. Authentic voices and sound effects are used to engage the visitor in a dramatic and more compelling experience. While the Alcatraz Guide uses rather aged technology with linear information access, the systems described in the following took a similar approach in preparing historic informations in a dramatic and narrative way, but tried to reduce the earlier mentioned drawbacks of linear access tours.

Dramatic and compelling experience

2.5 The Voices of Oakland

This audio-based augmented reality experience is situated in the Oakland Cemetery, USA. The basic concept is, that a virtual narrator lead the visitor through an overarching linear story and meanwhile guide her on a predefined route to various gravesites at which additional content could be selected [Dow et al. [2005]]. The cemetery is chosen as location for the project since not only its own history can be communicated, but also the buried people provide a starting point for giving information about important events that took place during the resident's lifetime. The stories and experiences associated with them serve to convey a more personal insight and a vivid and compelling experience.

Overarching linear story

Convey personal insight

The backbone of the whole experience is the continuous story line combining the different topics into one closed

Verbal navigational hints

narrative with a tension curve and a climax. The narrator is speaking directly to the visitor, leading her by navigational hints from one grave to the next (see figure2.4 a). The progress in the story is controlled by the visitor's position and a carried along controller (see figure 2.4 b), enabling to visit the site at a personal pace and to eventually take detours and watch graves not included in the tour.

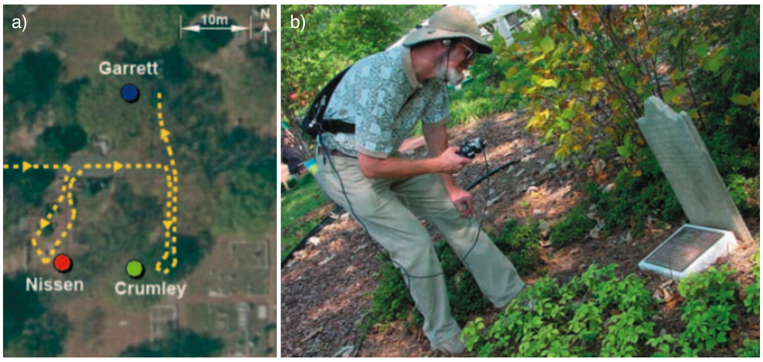


Figure 2.4: a) A satellite image of a part of the cemetery with circles at each grave and the intended experience path. b) A visitor wearing the system with the controller in his hands, reading a grave inscription (Both pictures taken from Dow et al. [2005]).

Select further contents

Arriving at a grave the narrator addressed the virtual resident, who then begins to unfold his story in the following conversation. Additionally, the visitor has, at every station, the ability to select further contents from three different categories: life stories, history, and architecture. To listen to an audio clip from a preferred category, the visitor has to press the category's button on the controller. When pressing the button of a category again, a different audio clip from the same category is played. Beside of this, the controller has buttons for pausing and rewinding the actual clip, to decrease and increase the volume, and to advise the narrator to lead to the next grave.

Wizard of Oz approach

The position depending parts of the experience, e.g., the navigation hints between the graves, are faked with a Wizard of Oz approach. Clips are manually selected and played, so if the visitor for example gets lost, she is helped to find back on the narrative path by simple navigation statements like “Go back ten steps” or “Turn around”.

The experience at the cemetery is, besides some selectable sub information, structured as a linear narrative, the following project tried to abandon this.

2.6 Riot! 1831

Like in the previous project the goal of Riot!1831 is to provide historic information in a dramatic and compelling way [Reid et al. [2005b]]. It lets the user experience the riots occurred on Queens Square, Bristol, England in 1831. Right at this place, it lets the user walk freely through an interactive audio landscape reproducing the happenings. The visitors can borrow a small back-pack containing a PDA with GPS receiver and headphones, with which they strolled around the 150 meter wide square. The area is divided into 34 regions (see figure 2.5), each associated with up to three different sound files. Moving into a region triggered one of the sound files playback, leaving the region caused it to stop. The closer a visitor gets into the centre of the region the louder the audio file becomes.

Interactive audio landscape of a riot

Devision in regions

The sounds are short vignettes based on the real events that took place at the square, for example, the visitor can hear the rioters' voices as they plundered the buildings surrounding the square, the flames as buildings burn, or merchants as they flee for their lives. The whole square is divided into four themed quadrants, one, e.g., reveals critical events by violent scenes with lots of action and fire, and another one contains more revelry with scenes of dancing and feasting rioters. Nevertheless, since the aim was a non-linear experience, the contained regions are nearly independent from each other and the clips associated to a region are mostly played in random order. In addition to that, a background sound file of a general crowd murmur is played in loop during the whole usage time to submit a background atmosphere and let the visitor infer whether the system is still running or not. Seventeen minutes after starting the experience a non located and everywhere hearable clip is played, which reflects the entering of a dragoon charge bloodily cutting down the rioters and the aftermath they left behind. In the mean time no other clips could be accessed. [Cater et al. [2005]].

Visitor hears rioters' voices

Non-linear experience

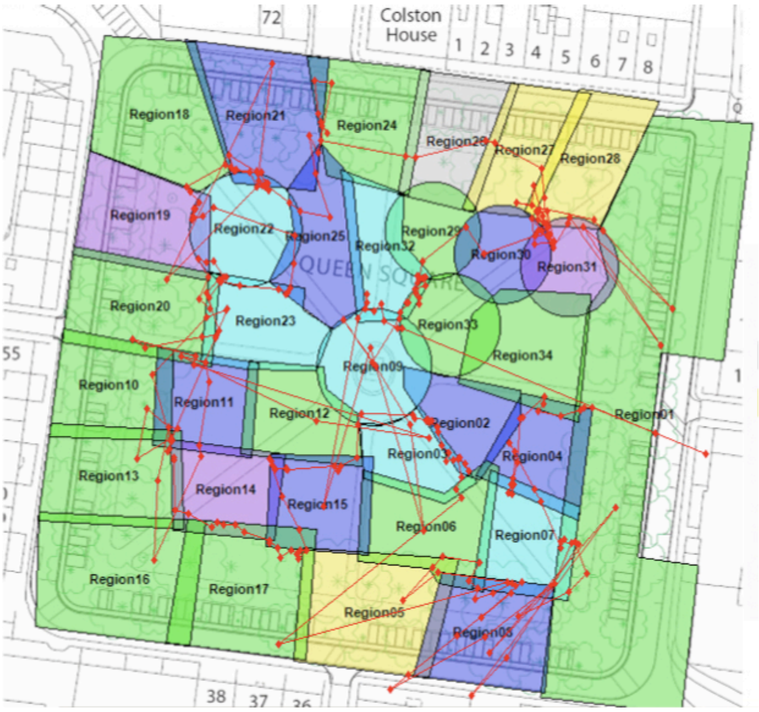


Figure 2.5: Map of the Queens Square overlaid with the region layout of the Riot! 1831 application (taken from Reid et al. [2005a]). The red line depicts the walked path of a visiting couple.

Prototyping and evaluation

The project was developed by the University of Bristol and the Hewlett-Packard Laboratories who provide the technological side and authoring tools, in collaboration with two artists writing the interactive play. The authoring tools evolved simultaneously with the progress of the project, incorporating the feedback of the writers who used it for prototyping and evaluation [Reid et al. [2004]]. Riot!1831 was brought to a general public in a three week trial period. During this time the experience was evaluated for over 700 visitors, resulting into over 500 hundred short questionnaires and movement log-files, 30 semi-structured interviews, and four in depth ethnographic case studies.

Feeling of confusion

The evaluation revealed, that many people expressed dissatisfaction with not being able to make enough sense of what is going on. A frequently reported feeling was that of confusion. The authors argued that the randomness in the sounds, which are encountered while walking around the square, is designed to emulate the confusion of a real riot [Reid et al. [2005a]]. Although in this context the idea might be appropriate, the concept is not always portable.

2.7 Navigation via Virtual Sound Sources

Navigated by audio signals

The main idea of the systems described in the following is to use a virtual sound source to help the user in finding a potential target position. With these systems the user is navigated by a signal which is presented through stereo headphones and modified to be received as coming from a specific direction. If the target is to the listener's right the signal is only presented on the right ear and when she turns her head to the left it fades over to the other ear.
The AudioGPS-System [Holland et al. [2002]] uses a short noise signal which is played in short intervals to represent the target. To communicate the distance to the target, the AudioGPS-System incorporates a geiger counter metaphor, which means, when the listener comes closer to the target the time intervals between the noise signals become smaller. The main aim of this system is to provide navigational aid without distracting the user from her environment.

AudioGPS-System with geiger counter metaphor

Guidance by music

A slightly different approach have the projects named "Navigation via Continuously Adapted Music" [Warren et al. [2005]] and "Melodious Walkabout" [Etter and Specht [2005]]. Both systems use a piece of music instead of a noise signal. To communicate the distance to the target, the volume of the music decreases with rising distance in the first project. In the second project an inverse mapping is used. The volume of the music remains constant until the user comes close to the target and the music is faded out.

Directional ambiguity

One challenge of this kind of systems is to enable the user to decide whether a source is in front or back of her (for a more detailed description of this problem and its origin see 4.2.3—"Problems of the Spherical Head Model"). Therefore some of the described systems add an additional cue into the audio signal to resolve this directional ambiguity. The

Incorporation of lowpass filtering

AudioGPS-System uses different noise signals for sources in the front and in the back of the listener. In case of "Melodious Walkabout" signals emitted by sources in the back of the listener are lowpass filtered. In case of the "Roaring Navigator"-Project, which we described earlier, sources in the back of the listener are completely faded out.

Chapter 3

First Interactive Protoype

In this section we will describe our first interactive prototype. We will mention which information we wanted to gain by implementing and testing it, and the resulting requirements the prototype has to fulfill. We will then give some insights into details of the implementation, and finally report about the evaluation and its results.

Goals of prototype

The main goal of our prototype, was to prove the concept of navigating only with the help of auditive clues through a virtual environment. First of all we wanted to know how easy it is to find the position of a source and how good the overall orientation is. Since all these points are strongly related to the quality of the spatial sound, this will also be an evaluation of the auralization. As a result we wanted to find out the impact of how the parameters of the rendering algorithm influences the audio experience and in which way they do.

3.1 Implementation

First of all we needed a possibility to assemble a small test scenario with several sound sources. It was our goal to test

Authoring environment

the influence of the different rendering parameters and layouts. Therefore a flexible scenario authoring environment, which lets us easily change those parameters and dynamically built up scenes would be helpful for our tests. Although being our target platform, the iPhone with iPhoneOS, is with its small display rather unpractical in this stage of the design phase. For that reason we decided not to implement our first prototype on a mobile device. Instead we chose Mac OS X as platform enabling us to use Objective-C. This brought up the advantage, that we could port the prototype to the iPhone more easily.

Mac OS X as platform

According to our aim that the prototype should be fully interactive, we needed a possibility to continuously track position and orientation of a listener in the physical environment. Since the tracking system which was originally planned to deliver the position in the final-product was not available on time, we used another tracking system available at our group. The Vicon high precision optical system uses several infrared cameras and spots to determine the position and orientation of objects which are marked by beacons.

Vicon tracking system

Furthermore, we needed an audio rendering framework responsible for the creation of the spatial sound. As our target platform is the iPhoneOs, there were only two possibilities to choose at that point: FMOD, a commercial framework, and OpenAL, an open source project. Both systems comply to the version 1.0 guidelines for interactive spatial sound defined by the 3D Audio Working Group of the Interactive Audio Special Interest Group, which declare a unique standard and a preferred implementation [3di [1998]]. The version of FMOD for the iPhone had, compared to OpenAL's iPhone implementation, slightly higher capabilities, e.g., basic signal filtering functions. Nevertheless, it is still a commercial software and to that date still in a developing state. OpenAL, in contrast, is standardly available on iPhone OS and Mac OS X by default. Furthermore, it is the recommended API for simultaneously playing multiple sounds on the iPhone. Therefore, we decided to use OpenAL whose structure and method of operation will briefly be explained in the next section.

Rendering Frameworks

3.1.1 OpenAL

OpenAL is a cross plattform API enabling the creation of three dimensional audio scenarios from prerecorded audio samples in realtime. It was developed as complement to the well known and spread OpenGL library. By following its pattern, OpenAL is similar in its API structure and in defining a unified interface hiding the actual implementation. As already mentioned, this provides an advantage for us, because we eventually can port our Mac OS X prototype code to the iPhone more easily.
With OpenAL we are able to define an audio scenario by positioning sound sources in a three dimensional space. In addition we can set position and orientation of the listener and OpenAL renders the spatial audio experience accordingly.

Cross platform API for spatial audio rendering

OpenAL offers three basic primitives for the creation of an audio landscape: source, buffer and listener. A source defines a point in space emitting a sound signal into every direction. Directed radiation is possible, but will not be used within this test. Beside the position parameter, the source has some additional properties influencing how the listener perceives the emitted signal, the *rolloff factor*, the *maximal distance*, and the *reference distance*, whose functionality will be explained later. To determine the emitted sound, a buffer, containing the actual audio data, can be attached to the source. The buffer holds the audio data in raw PCM format and is able to handle several sample sizes and rates as well as mono or stereo data. Finally, there is the listener object. It has a position, a viewing direction and an up-vector, according to which the spatialization algorithm synthesizes a signal.

Three basic primitives: source, buffer and listener

To get a realistic listening experience it is important that the source volume perceived by the listener decreases when the source is more distant. This is automatically done by OpenAL, but we have the opportunity to influence how this is done. Therefore six different models of distance attenuation and additional properties are provided. In the following we will describe the *Inverse Distance Clamped Model* (IDCM), which is recommended by [3di [1998]]. It creates a listening experience, which is closest to the real sound be-

Model of distance attenuation

havior compared to the other models, and is therefore used in our project. With respect to the source-to-listener distance, r, the $Gain_S(r)$ value is calculated by the following derivation steps, (3.1)-(3.3), and is used to scale the original audio signal to achieve a natural volume decrease (see figure 3.1).

Natural volume decrease

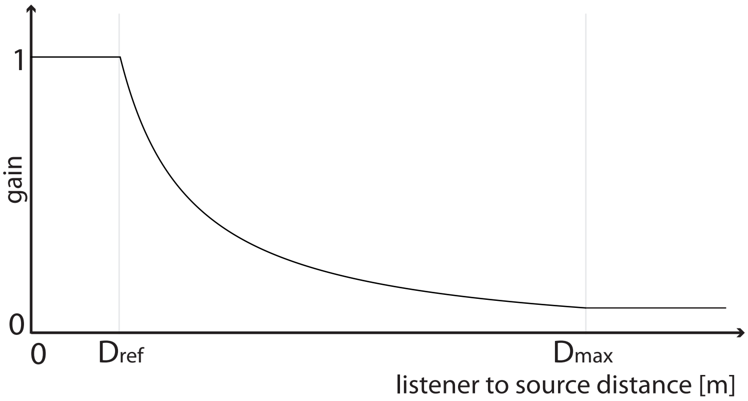


Figure 3.1: Source gain calculated with the Inverse Distance Rolloff Model.

$$r = \max(r, D_{ref}) \tag{3.1}$$

$$r = \min(r, D_{max}) \tag{3.2}$$

$$Gain_S(r) = \frac{D_{ref}}{(D_{ref} + R(r - D_{ref}))} \tag{3.3}$$

First we will look at (3.3). In case the distance r is equal to the reference distance D_{ref}, the term $r - D_{ref}$ is zero and the overall result becomes one. In this situation a listener perceives the signal which is emitted by the source at its original volume. If she now moves away from the source the volume decreases. The rolloff factor R controls how fast the source becomes quieter and lets us indirectly control how far a source is hearable.

Reference distance

Rolloff factor

If only (3.3) would be used, $Gain_S(r)$ would rise fast against infinite for distances smaller than the reference distance. So the distance r is limited in a first step to a lower bound given by the reference distance (3.1). By this the maximal gain value is one. Analogously r is limited by the maximal distance D_{max} with (3.2). This step is used to

Maximal gain value

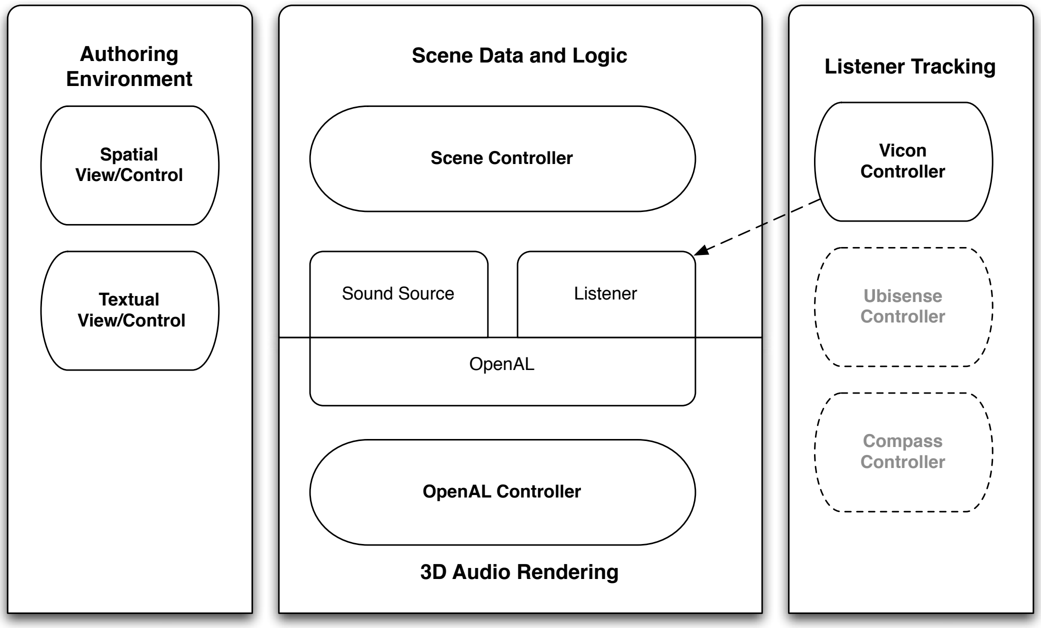


Figure 3.2: Conceptional structure of the prototype implementation.

set a lower bound for the volume attenuation by specifying a distance after which the volume does not decrease anymore.

3.1.2 Software Structure

Division into independent software parts

The software is structured into four parts, each one dedicated to different functionalities (see figure 3.2). The first part encompasses scene data and logic, the second a scene authoring interface, the third is dedicated to the control of the tracking systems, and the fourth enables the control of the audio rendering. The division into independent software parts facilitates an easy replacement of the different parts, e.g, to exchange the Vicon tracking by the other tracking systems. Additionally, the authoring interface is removable without restricting the use of the other parts, ensuring the portability of the prototype to the iPhone platform.
In the following we will briefly describe these different parts.

Scene data and logic

The scene data and logic part consists of scene objects wrapping the earlier described OpenAL primitives in an object oriented manner. There are a sound source object and a listener object. A sound source object encompasses a source primitive and a buffer primitive, therefore audio-files are directly assigned to a sound source object. Furthermore, a so called *Scene-Controller* enables the creation of scene objects during runtime, and the saving and loading of assembled scenes. The scene a user can create can consist of one listener object and potentially multiple sound source objects.

Tracking system control

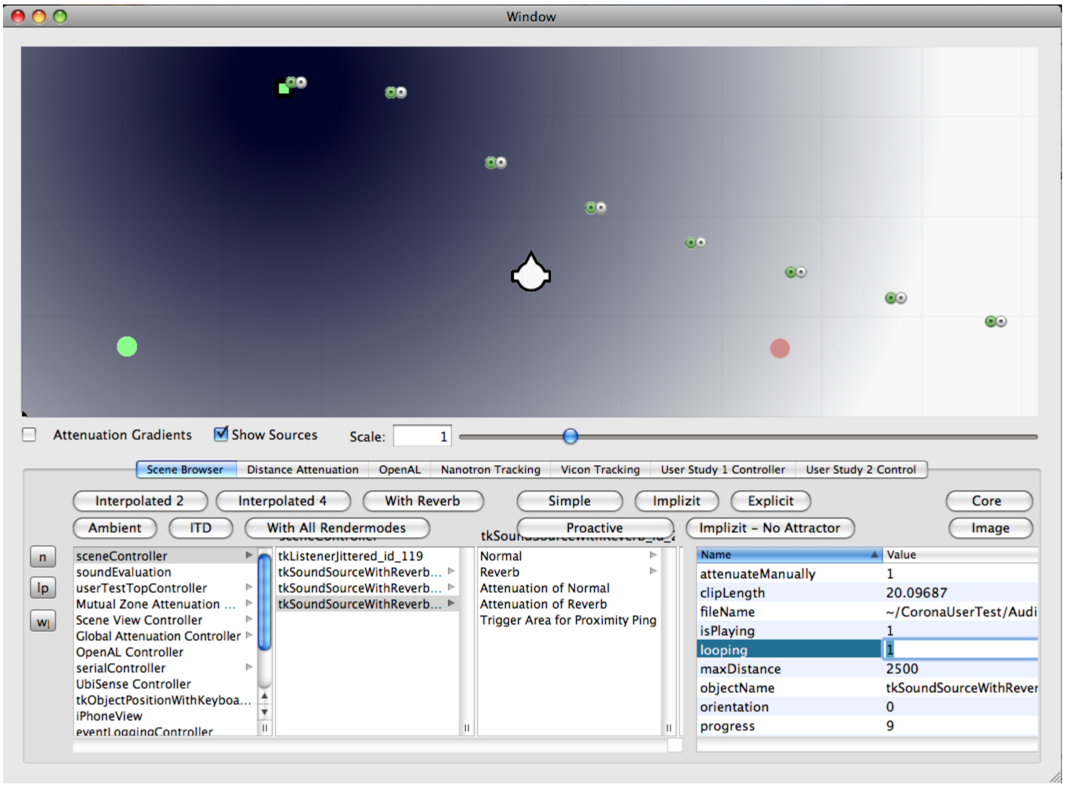
The tracking part for this prototype consists only of the so called *Vicon-Controller*. This controller establishes a network connection to the Vicon server and receives the listener's position and orientation. After that, it transforms the values from the Vicon coordinate space into the local scene coordinate space and updates the listener object. The Vicon system delivers tracking data with a much higher update rate and accuracy than our final tracking system. To be able to examine whether a poorer performance would potentially influence the listening experience, the controller has the additional functionality to reduce update rate and accuracy.

Audio rendering control

The most settings which influence the audio rendering, e.g., reference distance or rolloff factor, are controlled by scene objects which encapsulate the OpenAL- primitives. In consequence, the *OpenAL-Controller* is only responsible for a initialization of OpenAL and to enable the control of the used attenuation mode.

Authoring interface

The authoring interface consists of a spacial view/control of the scene, which offers direct manipulation of the position of the scene objects (see figure 3.3). It also visualizes the listener's orientation, the playing state of the sound sources, and if wanted the attenuation gradient for a sound source. The last-mentioned colorizes scene-view pixels according to the gain value of a source at the represented scene position. Thereby it helps, e.g., to get a quick overview about how far a source is hearable, and how strong audible ranges of different sources overlap.
Since not for every property a spatial view/control is useful, e.g., the audio-clip-filename of a source, we imple-



□ Attenuation Gradients ☑ Show Sources Scale: 1

Name	Value
attenuateManually	1
clipLength	20.09687
fileName	~/CoronaUserTest/Audi
isPlaying	1
looping	1
maxDistance	2500
objectName	tkSoundSourceWithRever
orientation	0
progress	9

Figure 3.3: Authoring interface of the prototype.

mented additionally a way to set parameters textually. In a tree-view, the before described controllers and scene objects, are hierarchically presented and reveal their properties on selection in a table-view for modification.

3.1.3 Audio Scenario

Sound source objects

One of the goals for this first interactive prototype was to prove that the concept of navigating between multiple dialogs works. Therefore the scene contains three sound source objects all emitting speech audio signals. (see figure 3.3). Its spatial dimensions are restricted by restrictions due to the Vicon tracking system.

Rendering attribute settings

We started with the following rendering attribute settings for all sound sources. The reference distance was set to one meter, the rolloff factor was set to one, and the maximal distance was set to one thousand meter such that it had no

influence in case of our scenario. The values were chosen since they should lead into a sound synthesis closest to the natural phenomenon.
This scenario was then used as starting point for the evaluation which will be described in the following.

3.2 Evaluation, Results and Discussion

Test of navigation-by-ear concept

The evaluation was done in an informal and experimental manner by the project members. To test the concept of navigation-by-ear, some of the members changed the positions of the sources into a spatial setup, unknown to the current listener, then the listener navigated through the scene solely by the help of auditive feedback. During the test we experimented with different rolloff-factors and reference distances.

High cognitive load

Several minutes of familiarization

Unnatural experience

Our tests showed that navigation solely by the help of the auditive feedback was only possible with a high cognitive load. One reason for this was, that to determine a source's direction the user has to turn his head continuously into strongly differing directions and highly concentrate on the changes in the perceived audio signal. Since this analytical procedure has to be found out and learned, it took several minutes of familiarization. Even then, the relative distances to the sources remained nearly undefinable, also for different rendering property settings. The perceived speeches sounded rather unnaturally through the absence of any reverberation. Altogether, there was no feeling of moving through a spacial audio scenario with speech signals emitted from several specific positions. The mapping between the listener's movements and the changes of the perceived signal felt sometimes quite arbitrary and there was only little comprehension of the spatial source setup.

Concept of navigation-by-ear not proven

In summary we can say that the navigation solely by the auditive feedback was not possible without a high cognitive load and several minutes of learning. Therefore, it could not be proven, that the concept of navigation-by-ear and a pure audio experience works. Since we believe that the fail of the concept was due to the bad spatialization of the sound sources, the main conclusion was that we have to improve our audio rendering. Therefore the next chapter is dedicated to the topic of spatial sound synthesis.

Improve audio rendering

Besides that, the possibility to change the test setup dynamically, showed to be beneficial for the informal and experimental test method, because it allowed a rapid scene prototyping cycle.

Chapter 4

Foundations of Auralization

In this chapter we will explain the basics of sound propagation and spatial hearing. We will only focus on the aspects we need in the context of our application.

AURALIZATION:
Auralization is the process of rendering audible, by physical or mathematical modeling, the sound field of a source in a space, in such a way as to simulate the binaural listening experience at a given position in the modeled space [Vorländer [2007]].

Definition:
Auralization

4.1 From the Sound Source to the Ear

To understand the phenomenon of sound we will start with one of its origins: the movement of a physical object. A moving object hits particles of the embedding medium, e.g., air, shifts neighboring particles and induces a difference in the local air pressure from the normal one. This deviation propagates as a wave phenomenon in the medium. A wave is a disturbance of the equilibrium state traveling through time and space.

Difference in local air pressure

It can be described in a general form by

$$u(x,t) = u_0(x \pm ct). \tag{4.1}$$

Sound pressure level

The equation (4.1) describes that a disturbance at time-point $t = 0$, described by $u_0(x)$, moves in the time-period t with speed c the distance vt in x-direction. (see figure 4.1). In case of airborne sound waves this disturbance is a local deviation from the normal atmospheric pressure and is called *sound pressure level*. The velocity c of the wave is also given by the travelled medium and can be approximated by $c = 343\frac{m}{s}$, under conditions of 20° C in dry air. Not the physical particles propagate in a wave – they only move small distances around their equilibrium state – but the energy which is stored in them as potential and kinetic energy. Since the emitted energy at one point in time spread uniformly in every direction from the source with c (we will only look at omnidirectional monopole sources here), it is after time t_1 distributed over the surface of a sphere with radius $r = t_1 * c$. Thereby, if we assume a constant energy emission per time unit, P_{ak}, and want to know the wave's intensity, I, floating through a surface part with source distance, r, we have to solve the following equation:

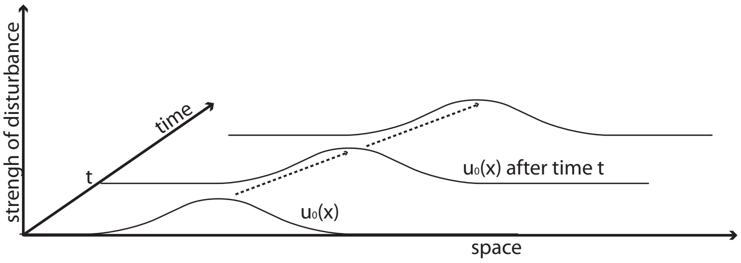


Figure 4.1: A disturbance $u_0(x)$ traveling through time and space.

Spread in every direction

$$I(r) = P_{ak}\frac{1}{4\pi rr} \tag{4.2}$$

Threshold of hearing and pain

The sound intensity is the main factor in the perceived loudness of a sound. The range of sound intensity humans can perceive starts with the threshold of hearing, $I_{TOH} = 10^{-12}\frac{W}{m^2}$, and ends with the threshold of pain, $1\frac{W}{m^2}$. To describe this very large range often a logarithmic scale is used to describe the sound intensity level L_W, expressing the intensity with respect to the threshold of pain:

$$L_W(I) = 10 \log_{10}(\frac{I}{I_{TOH}}). \tag{4.3}$$

Distance law of intensity

By adapting and rewriting the equation we obtain the *distance law of intensity* which enables us to derive the intensity level L of a wave signal emitted at a distance r with an intensity level L_W:

$$L(r) = L_W - 20 \log(r) - 11. \tag{4.4}$$

Although this equation looks very unrelated to the equation used by OpenAL (see 3.1.1—"OpenAL") to simulate distance attenuation both roughly lead into a level reduction of approximately $6dB$ per distance doubling.

Waves distorted by obstacles

Up to now we only talked about a free-field situation with no objects influencing the propagation of the sound wave. If this changes and waves hit obstacles, they are distorted by reflection, diffraction and scattering, inducing, e.g., echoes in a room and a reverberant environment. But before a sound can be perceived, another obstacle –the listener herself– is hit and modifies the sound wave. In the following section we will describe how the anatomy of the listener influences the sound waves and how the listener can use this to get a spatial impression.

Listener as obstacle

4.2 Binaural Cues

Left and right ear signal distorted differently

Since the human ears are spatially separated, pointing at different directions and having the head as a physical obstacle in between them, a sound has to travel along different paths until it reaches each ear. Being exposed to different physical effects, the signals get distorted differently on their respective path, allowing a spatial perception. In the following we will briefly describe the differences in the left and right ear-signal which are important for source localization and how humans use these differences to localize a source.

4.2.1 Interaural Level Difference

Level of diffraction depends on wavelength

The interaural level difference (ILD) is the first discovered and investigated disparity between the signals reaching the left and right inner-ear of a listener [Strutt [1907]]. It describes the intensity differences in the left and right ear-signals caused by collision of the sound waves with the listener′s body. Since the level of diffraction of a sound wave impacting a physical object depends on its wavelength, the same holds for its ILD [Hartmann [1999]]. In figure 4.2, we can see the dependency between the ILD and the wave frequency. It is calculated for planar waves and a listener which is modeled by a sphere with opposite poles as ears. Using such a simplistic head model, we can see that the ILD function of frequency and source azimuth is already very complex.

ILD decrease below 1000Hz

It is remarkable, that the ILD rapidly decreases for frequencies lower than 1000 Hz. The reason for this is a decreasing diffraction of sound waves, if their wavelength is longer than the diameter of the head. For example, in case of a sinus signal with a frequency of 500 Hz the wavelength is, with a value of 70 cm, more than four times the average head diameter, leading to an ILD tending to a range not perceivable for humans.

ILD dependency for short distances

The function describing the ILD even gets more complex if the impacting sound waves can not be considered as planar anymore, which is the case if the distance between listener and source is smaller than one meter [Shinn-Cunningham [2000]]. In consequence, the wave diffractions around the listener, and with it, the distortion of signals emitted in the near field does not only depend on direction and position, but also on the relative source-listener-distance. In figure 4.3 we can see the ILD dependency of source distance under conditions of different directions and for a specific frequency. We can notice more than a doubling of the ILD if we change from the far to the near field in case of a sound source at 90^o azimuth.

Beside the ILD functions calculated for a spherical head, we can see in figure 4.3 some empirical ILD values. The measurements are made with microphones at the ear canals

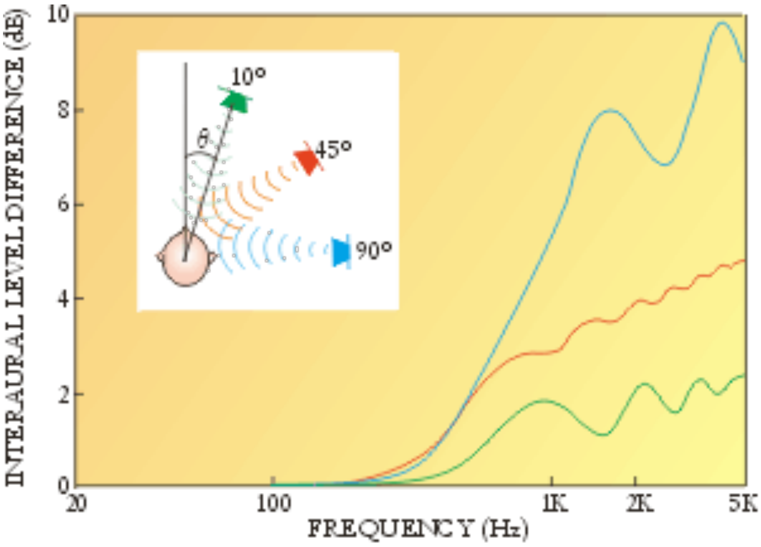


Figure 4.2: The ILD in dependency of the wave frequency for different azimuths (taken from Hartmann [1999]).

of a human subject. In case of 0^o azimuth we can observe that the measured ILDs are not zero. The reason for this is that a real head is (in most cases) asymmetric and therefore creates a dichotic listening, even if a source is located in the median plane [Shinn-Cunningham [2000]]. Thus, the spherical head model lacks some realism which may cause some negative effects we will describe in 4.2.3—"Problems of the Spherical Head Model".

Lack of realism

Nevertheless, the prediction of the spherical-head model, that the ILD decrements in case of long wavelengths into a not perceivable range, was proved by empirical measurements. Since we are able to localize the origin of signals containing only frequencies below 500 Hz, there must be another cue coded in the dichotic signal.

4.2.2 Interaural Time Difference

Looking at the spherical-head model again, we can notice that in cases of azimuths, which are different than 0^o and 180^o, the path-lengths which a sound wave has to travel are different for the left and the right ear. In figure 4.4 we

Path-lengths differ for left and right ear

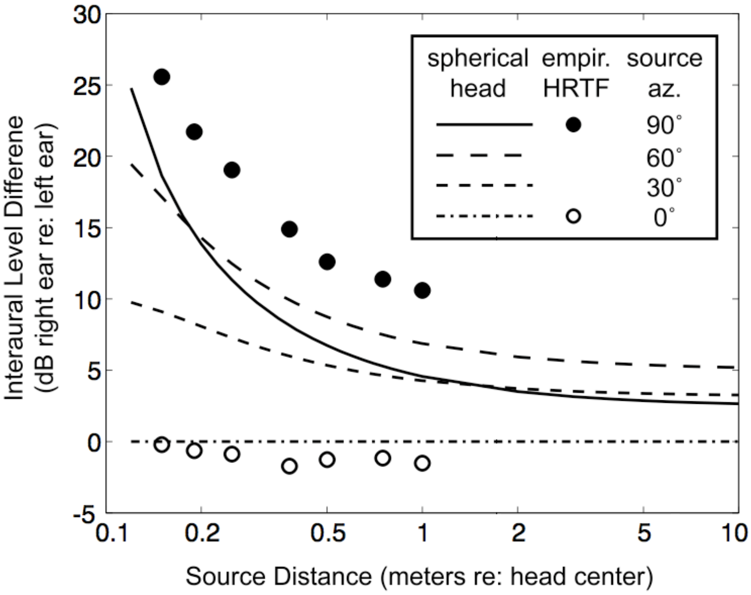


Figure 4.3: The ILD in dependency of the source to listener distance for a specific frequency (taken from Shinn-Cunningham [2000]).

see a planar sound wave striking the spherical model of a head. It arrives with speed c from a direction specified by the azimuth angle θ. The traveled distance between left ear signal and right ear signal differs by:

$$a\theta + a \cdot sin\theta, -90^o \leq \theta \leq 90^o. \tag{4.5}$$

And the according interaural time difference can be calculated by:

$$ITD = \frac{a}{c}(\theta + sin\theta), -90^o \leq \theta \leq 90^o. \tag{4.6}$$

With an average head diameter $a = 8,75cm$ and a sonic speed of $c = 343\frac{m}{s}$ in air at room temperature, the values of the ITDs range from $0ms$, for sources in front of the listener, to about $0.7ms$, for sources with 90^o or 270^o azimuth. Tests have proven that humans can use the ITD as clue to estimate a sound source position.

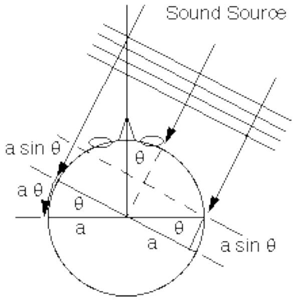


Figure 4.4: A planar wave hitting the spherical model of a head (taken from Duda [2000]).

4.2.3 Problems of the Spherical Head Model

We got to know the ILD and ITD which are the main cues for source localization, and used the spherical head model to calculate how a signal will be shaped when hitting the listener. One would assume now, if a signal is shaped according to this model and presented to a listener through headphones, that she would perceive the signal as a real sound source at a predefined position. Unfortunately this is not always the case. If the listener tries to locate the position of the source, she will encounter some problems.

Front-back-confusion

A first one is known as *front-back-confusion*, and is caused by the symmetric geometry of the model. This symmetry leads to the same ILDs and ITDs for sources in directions having an equal angle distance to the ear-axis (see figure 5.3). The directions with the same ITDs and ILDs form the so called *cone of confusion*. If a listener hears a sound source, e.g., a dialog between two persons, she cannot decide where on the cone of confusion the sound source is, respectively whether the persons are in her front or back [Blauert [1983]].

Cone of confusion

In-head-localization

A second problem is that of *in-head-localization*, which describes the listeners perception of a sound source location between her ears. This phenomenon is caused by small abnormalities in the artificial signal shaping differing from the shaping in a real scenario [Durlach et al. [1992]].
Both problems are rooted in a too strong simplification of

the listeners anatomy by the spherical head model. A more accurate description of occurring signal distortions can be explained by the anatomical correct transfer function which will be the topic of the next section.

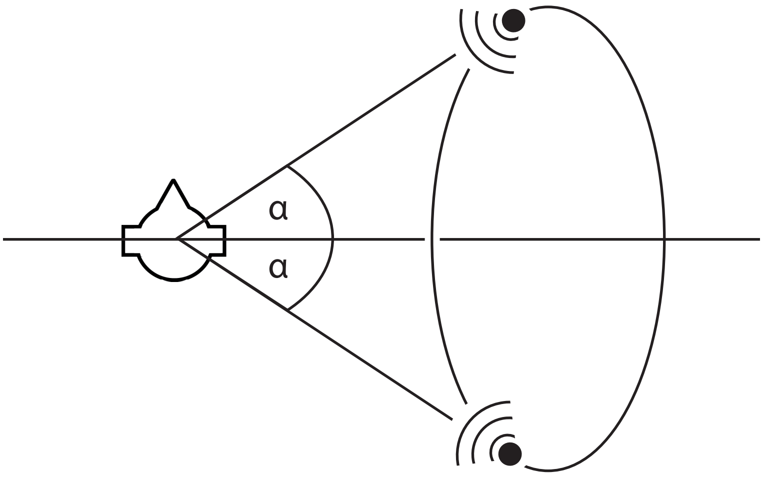


Figure 4.5: Two sound sources with the same ear axis angle sharing the same cone of confusion.

4.2.4 Anatomically Correct Transfer Function

Head-related transfer function (HRTF)

Dependency on listener's anatomy

As already mentioned at the beginning of this chapter a sound wave hitting the listener gets disturbed by reflections and diffractions at the head, the torso, the shoulders, and the pinnae. The head-related transfer function (HRTF) is a complete and formal way of describing these distortions. It is defined as the ratio between measured sound pressure at the ear channel entrance and the sound pressure measured at the same position without the head [Vorländer [2007]]. The result is a complex response function which is different for every direction of sound incidence and depends on the physical anatomy of the listener. In case of a source distance below approximately one meter the HRTF also depends on the distance. The reason for that is, that the curvature of the wavefront is no more negligible [Shinn-Cunningham [2000]].

Using an HRTF in auralizing a signal coming from a spatial source in a desired direction, we first have to measure the HRTF for this direction. If we then convolve an arbitrary signal with the recorded HRTF, it sounds for the listener as if the signal is emitted by a real sound source in the chosen direction.

Auralizing a signal

If we want to be free in choosing the relative position of the sound source to the listener, as it is necessary when the listener should be able to move freely around the source, we have to measure an HRTF for every possible direction. Because this would be too time consuming in some situations, the HRTF is only measured for discrete sampling of the hemisphere. If we want to simulate a direction between two samples, the HRTF of the next two sampling points are linearly interpolated.

Measure HRTF for every direction

The HRTFs of different individuals can differ strongly from person to person. The main reason of the differences are strong variations of pinnae shapes, which lead to several resonances and antiresonances. Presumably through a lifelong experience a person learns how to use the given cues. In a context where it is unsuitable to measure the HRTF for every listener it is common practice to use a function set which was proven to work well for a large user population. Another possibility is to create an artificial head model with an average anatomy and use this for the measurements [Vorländer [2007]].

Strong variations of pinnae shapes

The question is now, whether the usage of a non-individualized HRTF brings back the earlier described problems of *in-head-localization, front-back confusion* or even worse effects. The answers to this question given in the literature are quite different.

Non-individualized HRTFs create problems

In [Zotkin et al. [2002a]] it is said that the HRTFs "are not interchangeable" for different persons. In [Vorländer [2007]] it is stated that using HRTF, measured on a basis of a standard-dummy-head, leads into strong disturbances in the listening experience. In [Wenzel et al. [1993a]] a test is described in which a virtual sound source scenario is auralized with non-individual HRTFs. In result the 16 of 18 subjects perceived no difference between this setup and an identical scenario with real sound sources. Another test made by the authors of [Wenzel et al. [1993a]] showed that the front-back-confusion was increased, when using non-individual instead of individual HRTFs. In summary we can say that when using non-individual HRTFs a good localization experience is not guaranteed.

4.3 Binaural Hearing at a Cocktail Party

Looking at our planned target scenario, we notice that in some situations the listener hears multiple sources at once. To make sense of the speech signals which are contained in the overall perceived signal, the listener has to segregate the different signals from each other. This problem relates to the *Cocktail Party Effect* which describes the listeners capability to focus her listening attention on a single talker among a mix of conversations and background noises [Cherry [1953]].

Cocktail Party Effect

It has been shown that the method which is used in the auralization process has an influence on how good a listener can segregate and understand a speech signals while she is exposed to one or more competing sound sources [Hawley et al. [2004]]. The influence of the used method originates from two effects caused by a binaural presentation of audio signals: *energetic unmasking* and *binaural unmasking*..

Energetic and binaural unmasking

To understand the origin of the energetic unmasking and in which cases it occurs, we take a look at a small example: Let us consider a scenario with three talking speakers positioned in front of a listener. In consequence all speech signals arrive with equal volume at both ears and the intelligibility of every speaker is equally worse. Now two of the speakers, the maskers, walk off to the right side of the listener. Due to the head shadow and the induced ILD, the signals of the maskers have an about $16dB$ lower level on the listener's left ear than the signal of the speaker who is in front of her. By this the listener has access to a cleaner signal of the frontal speaker, which rises it intelligibility [Hawley et al. [2004]].

Access to cleaner signal

Now one of the maskers on the right walks over to the listener's left side. In this setup the listener has no ear with a reduced masker level and there is no intelligibility advantage by energetic effects.

A second advantage for speech segregation and intelligibility is due to the differences in interaural time delay between competing sources. The delay causes a binaural unmasking of the low-frequency parts of the speech signal and thereby increases intelligibility. Hawley et al. [2004] showed that this increase is not dependent on the maskers being in one hemisphere of the listener. The authors also showed that these advantages are robust even in complex hearing scenarios with up to three speakers.

Increased intelligibility

Until now we mainly talked about how the perceived directionality of a heard signal evolves and how we can artificially rebuild it. To know the actual location of a sound source the listener also needs to know its distance. Therefore this will be the topic of the next section.

4.4 Distance Cues

This section will handle all relevant parts needed to understand distance perception and how this knowledge can be used in auralization. To answer the question how a listener can estimate the distance of a sound source, we take a look on how the perceived signal differs for different source distances. After that we will describe how they contribute to the listener's perception of distance.

4.4.1 Perceived Signal

In case the listener-source distance is larger than one meter, the intensity of the sound wave perceived by the listener is reduced by $6dB$ per every doubling of the distance due to the earlier described reasons (see 4.1—"From the Sound Source to the Ear"). If the distance is larger than approximately 15 meters the influence of the travelled medium becomes relevant. Since mostly the high frequencies are absorbed, the sound becomes a more and more muffled character with rising distance. If the source is closer than

Three distinct distance cases

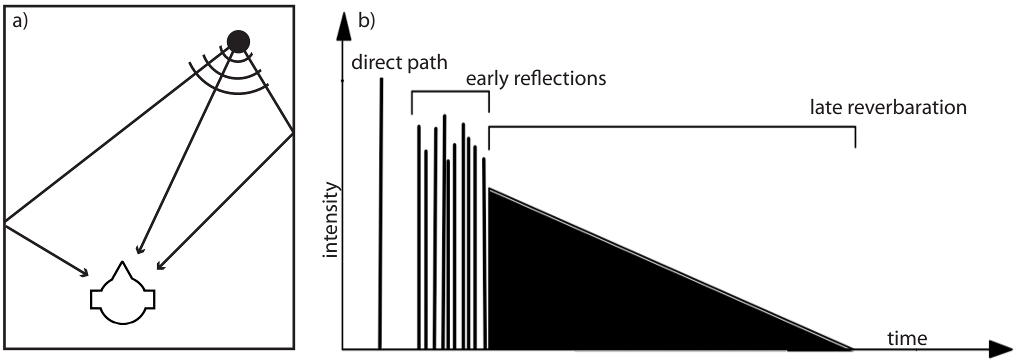


Figure 4.6: a) Reflection of sound waves in a room. b) Impulse response of a room.

Near-field-hearing

approximately one meter, the already mentioned growing curvature of wave fronts is relevant. In this situation, often named *near-field-hearing*, the source direction also influences how fast the level changes with the distance [Blauert [1983]].

Exclude two cases

Taking into account the sound source setup in our planned scenario, we can already leave out two of the described effects. Since sound sources will be arranged in a way, that only sources are perceivable which are nearer than 15 meters, the long distance effect will not occur and is therefore negligible. The same applies for the small distance effect, as we will handle the area directly around the sources in a special way due to reasons of usability and restrictions by the used position tracking system.

Physical objects in environment

Multiple sound paths to listener

Until now we did not regard physical objects in the environment beside the sources and the listener. Adding more objects to the rendering scenario complicates the auralization process, even if the objects do not emit sound. As we can see in figure 4.6a, the sound is reflected from walls, the floor, the ceiling, or other objects. This is why the sound travels on multiple paths from the source to the listener. This indirect paths depend on the scene setup and are different for every listener position. One can notice that the paths have different lengths and hit different obstacles. In consequence, the signals reache the ears with different delays, intensities and spectral shapes [Vorländer [2007]].

In figure 4.6b we can see a schematic depiction of an impulse response for a room and a specific listener position. Every peak represents one indirect path and we can read out its intensity and delay. The response can be divided into three parts: the direct path, the early reflections, and the late reverberation tail. The signals encompassed by the first two parts are still distinguishable from each other by the listener and depend on her position. The late reverberation only consist of a dense succession of echoes which decay exponentially. It has a diffuse sensation and can be seen as to be independent of the listener's position.

Reverberant environment

Since low frequency waves diffuse more effectively around obstacles in the direct and indirect sound paths, the frequency spectrum of the signals changes with the listener's position in the environment [Brungart [1998]].

Low frequency waves diffuse more effectively

4.4.2 Perceived Distance

As the pressure level decreases with $6dB$ for every doubling of the distance, the intensity of the perceived signal seems to be the most obvious cue for distance estimations. But in case the listener does not know the original intensity of the source, e.g, at the distance of one meter, she is not able to use it as an absolute cue for the distance [Mion et al. [2007]]. In case that the listener has knowledge of the intensity for a reference distance, e.g., if she has heard the source from a different position, she is theoretically able to estimate the actual distance. One may assume that the listener unconsciously uses the described distance law to estimates sources distances. Nevertheless this is not always the case. In Begault [1991], the authors describe that 70 percent of their tested listeners preferred an intensity decrease of 9 dB to have the sensation of a half of the source distance. The tests were conducted with speech signals.

No absolute cue

Intensity decrease of 9dB preferred

A reason for this could be that in case of familiar signal types, like, e.g., speech signals, the intensity sensation is cognitively associated with a typical distance [Gardner [1969]]. Nevertheless, this associations need not correspond to the physical law.

High frequency attenuation increases perceived distance

Another distant hint which is used by the listener are spectral distortions caused by the absorption of the high frequencies in the air and the diffractions at obstacles. It has been shown, that the attenuation of the high frequencies increases the perceived distance [Brungart [1998]]. Since this perception is independent from the underlying physical phenomenon, there is no need that the signal shaping, e.g., an increasing lowpass for a rising distance, is based on physical correct derivations.
Nevertheless, this kind of distance cue has several disadvantages: Tests have shown that the perceived distance differs between different listeners; The reduction of high frequencies could decrease the quality of a signal and thereby, e.g., the intelligibility of speech signals; When the unshaped source signal is unfamiliar to the listener the spectral shaping is only a relative distance cue.

Absolute distance cue

Reverberation reduces cognitive load

The latter is different for the next distance cue. Since the intensity of the reverberation is nearly independent of the listener's position, but the intensity of the direct path signal increases with rising distance, the ratio between these two values is used as absolute distance cue [Blauert [1983]]. Also, it is said in Mion et al. [2007], that adding reverberation leads to a spatiality which changes the distance perception from an analytical process by loudness inference, to a more familiar everyday listening experience and thereby reduces the cognitive load.

Motion of the listener

Until now, we disregarded the motion of the listener. But while passing a source the perceived azimuth of the source changes with every step. When the listener walks with constant speed, she can use the angular velocity of the azimuth to estimate the source distance. In theory the listener could infer the source distance by

$$D = \frac{S \sin(\alpha_1)}{\sin(\alpha_2 - \sin(\alpha_1)} \tag{4.7}$$

More precise distance estimations

with the walked path length S, the first azimuth α_1, and the second azimuth α_2 (see figure 4.7). In practice this distance hint has showed to help the listener on making more precise distance estimations [Speigle and Loomis [1993]].
The reliability of these cues suffers if the listener is not sure whether a source has a fixed position or not. In figure 4.7

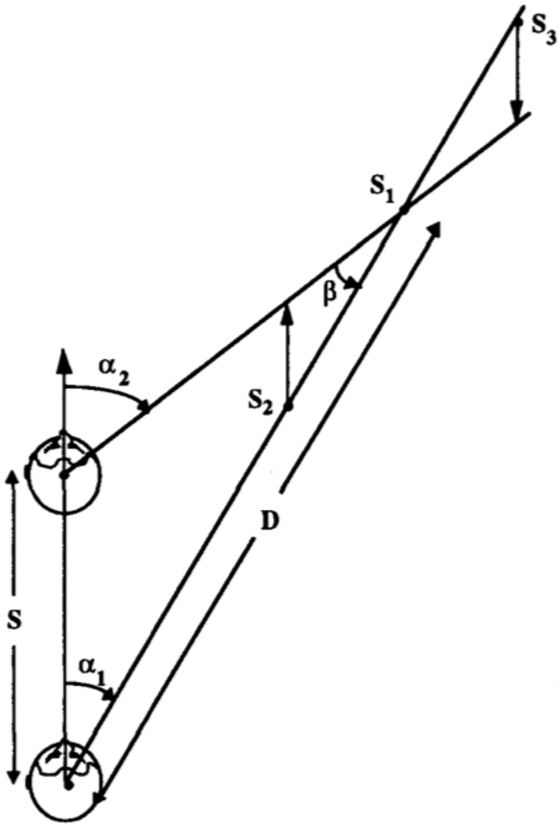


Figure 4.7: Stationary source S1 undergoes a change in azimuth from α_1 to α_2 as the listener moves through a distance S (taken from Speigle and Loomis [1993]).

Perception of moving source

we can see a stationary source with position S_1 and a listener passing it. She perceives the correct azimuth angles α_1 and α_1 but estimates the source to be at position S_2. In consequence the listener will believe that the source slowly moves in the same direction on a path parallel to her own. Analogously, if she estimates that the source has a position S_3 with a greater distance, the listener will perceive the source to move in the opposite direction. Therefore it can be beneficial in case of a static source scenario to communicate to the listener that sources do not move.

Chapter 5

Auralization

Now that we have a better understanding of the phenomenon of sound propagation and spatial hearing, the question arises how we could use this to improve the auralization in our project. In this chapter we will describe the implementation of three auralization methods which incorporate directional cues in different ways. Furthermore, we will explain how we integrate a simple room acoustic. In the end of the chapter we will describe a preliminary user study and the incorporation of the results.

5.1 Implementation

Usage of offline filtered audio signals

Since we are still bound to the OpenAL API and the fact that the underlying implementation on the iPhone is not capable of any kind of signal filtering, we have to find another way to bring the directional cues which we described in the last chapter, into the signal perceived by the listener. In the following we will describe how we use different offline filtered audio signals to bypass the lack of convolution possibilities.

5.1.1 Method A - Head-related transfer function

Discrete sampling of listener's hemisphere

The first method we implement incorporates signals which are filtered with an HRTF to create the spatial cues. As described in 4.2.4—"Anatomically Correct Transfer Function", it is the standard procedure when using HRTFs in auralization to use a limited number of HRTF samples measured for a discrete sampling of the listener's hemisphere. To auralize signals for directions which are not included by the sampling, the measured HRTF of the next two samples are interpolated and then convolved with the signals. Since we lack the possibility of online convolution, we have to do this processing step offline and to interpolate between the already convolved stereo audio signals at runtime.

Different order of interpolation and convolution

Since we use a different order in which the interpolation and the convolution are performed, the resulting signals of our method differs from signals created with the standard procedure. Therefore we will have to evaluate whether the approach leads into a satisfactory localization performance or if disturbing audio artifacts occur. Nevertheless a related procedure is described and evaluated in Algazi and Duda [2005] with positive results.

In a first step to implement this method we sample the whole range of possible azimuths $]0..2\pi]$ uniformly with a discrete set of values

$$\Gamma = \{\gamma_1, \gamma_2, .., \gamma_n\} \tag{5.1}$$

Sub audio clips

where n is the number of taken samples. In addition we create n sub audio clips for the audio clip of every source S which is to be auralized:

$$s_{\gamma_1}, s_{\gamma_2}, .., s_{\gamma_n} \text{ , with } \gamma_i. \in \Gamma \tag{5.2}$$

Each clip s_{γ_i} contains the original clip of S convolved with an HRTF corresponding to a specific source azimuth α_i. In the auralization process after every movement of the listener we calculate the current azimuth β of the source S with respect to the current listener's position and head orientation (see figure 5.1). Afterwards, we select the two γ_i

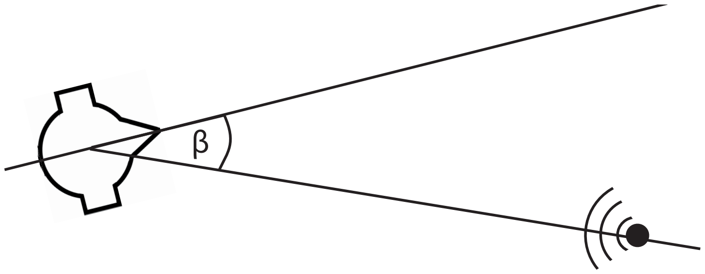


Figure 5.1: Relative source to head orientation, β.

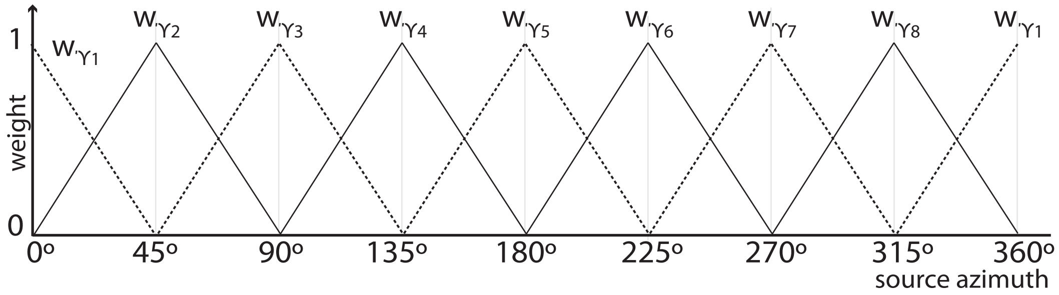


Figure 5.2: Weights of the different sub signals.

out of Γ which are closest to β. Thereby we obtain γ_a as the closest and γ_b as the second closest azimuth sample with respect to the current actual azimuth. The audio clips s_{γ_a} and s_{γ_b} therefore contain the audio signal with the binaural cues coming closest to the cues which would be heard in reality. To create the actual spatialized signal for the source S the audio clips s_{γ_a} and s_{γ_b} are added each with a specific weight $w_{\gamma_a}, w_{\gamma_a}$. The weight $w_{\gamma_{i_a}}$ of clip s_{γ_a} depends on how close γ_a is to β and the number of samplings which are used:

Sub audio signals added with specific weights

$$w_{\gamma_a} = |(\gamma_a - \beta)| \frac{n}{2\pi} \tag{5.3}$$

In figure 5.2 we see the contributions of each audio clip s_{γ_i} to the overall signal of S in dependency of the relative source azimuth β in case of $n = 8$.

Since the relative source azimuth β depends on the actual orientation of the listener's head, and thus changes several times a second, all audio clips s_{γ_i} of a currently auralized source have to be buffered in the working memory. Due to the hardware limitations of the iPhone, especially the

Limitations of iPhone

size of working memory, we selected the sampling density of the hemisphere to 45° steps. Our density is much lower than the normal used densities of about 3° [Vorländer [2007]]. Nevertheless, early internal tests showed that the listening experience and localization performance is improved against our initial test. The authors of Algazi and Duda [2005] used a comparable density and also gained positive results.

Low sampling density

As already mentioned in 4.2.4—“Anatomically Correct Transfer Function”, the listening experience and localization performance can be reduced if HRTFs are applied which do not correspond to the listeners' anatomy. Nevertheless, using individual HRTFs is unsuitable for our project, so we have to use a generalized set [KEMAR[1]].

Generalized HRTF

For the above mentioned reason and because we use a low sampling density, a sufficient performance of this auralization procedure for our purposes is not guaranteed. Nevertheless, the question is whether our application requires the listener to extract “realistic” three-dimensional information from the presented audio signal, or whether it is sufficient that the listener is able to approach a sound source. In the last case, a coarse simulation of ITD and ILD cues may be sufficient [Stanney [2002], Loomis et al. [1991]].

Sufficient performance is not guaranteed

5.1.2 Method B - Frequency independent ILD and monaural simulation of the pinnae shadow

This auralization method delegates the creation of the binaural cues to OpenAL but incorporates an additional monaural cue to decrease the problem of front-back confusion (see 4.2.3—“Problems of the Spherical Head Model”). The implementation of OpenAL on the iPhone uses a simple stereo panning of a source signal to the left or right ear to spatialize a signal. The panning thereby only depends on the deviation of the source direction from the listener's median plane (the plane between the left and right hemisphere of the listener). Thus, the created cues are symmetric for the front and back hemisphere of the listener. As described in

Additional monaural cue

[1] http://sound.media.mit.edu/resources/KEMAR.html

4.2.3, symmetric binaural cues create a cone of confusion and thereby lead to the problem of front-back-confusion. The main idea behind this auralization method is to reduce this directional ambiguity by incorporating a lowpass filter for signals from sources behind the listener. This idea was already tested with positive results in, e.g., Loomis et al. [1991] and Etter and Specht [2005], .

Main idea: reduce directional ambiguity

In addition to the front-back-confusion, the cone of confusion leads to a second problem when a scenario contains multiple sound sources. In fig 5.3 we see a scenario with two actually separated sound sources which have the same angle to the ear axis. Both sources are perceived by the listener with the same ILD. Reconsidering the earlier described cocktail party effect, this causes a cancelation of the advantages of any energetic effects. The consequence is a reduced intelligibility of both emitted signals and a reduced ability to selectively pay attention to each of the audio signals (see 4.3—"Binaural Hearing at a Cocktail Party").

Problem: Reduced intelligibility

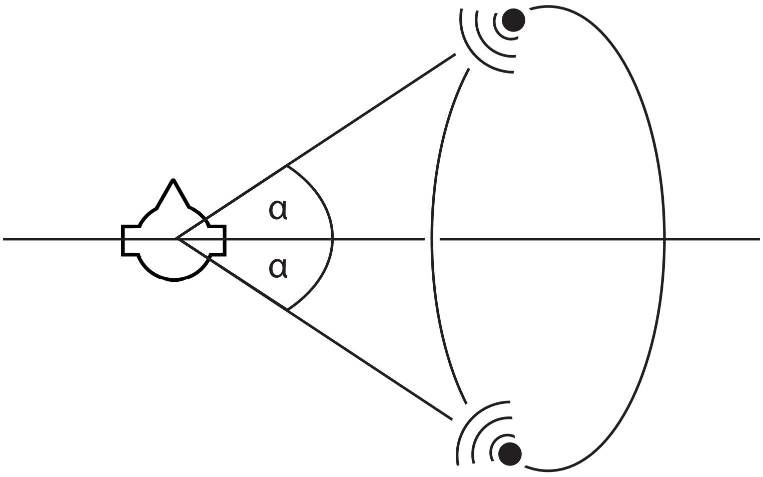


Figure 5.3: Two sound sources with the same ear axis angle sharing the same cone of confusion.

Spieth et al. [1954] try to resolve a related problem. In several experiments they tried to find conditions which enable a communication operator to identify best an important voice message out of several simultaneously played irrelevant messages and to perceive its content. The message

Related problem was investigated

contained a keyword and a following trivial question. The keyword identified the message as the important one and the question had to be answered by the operator to check whether he was able to follow the message. We think, that this task is closely related to our problem. Since the element of overhearing a keyword and listening to the message afterwards maps to the ability of the visitor to identify an interesting sound source in multiple simultaneously heard sources, and afterwards to attend only to the interesting source.

Identify and listen to important message of several

The results of the experiments showed that the ability to identify an important message was 15 percent higher when the signal of the message was shaped with a lowpass filter, no matter whether the irrelevant or the interesting message was filtered. Nearly the same enhancements hold for the task of following the content of the important message after identifying it. Another tested condition showed that, when the aural shaping is combined with a spatial separation of the sources, no negative effects occurred, with respect to the tested qualities.

Benefit from lowpass filtering

Workaround for filtering

Since OpenAL is not capable of applying a lowpass filter online, we have to find a workaround again. We use two sub audio clips for every audio clip s which we want to filter. One with the original audio clip s_{orig} and one with an offline filtered version s_{low}. The latter is created with a lowpass filter cutting off frequencies above 1000 Hz with an external software. During runtime we play both clips simultaneously. To simulate the filtering we crossfade between s_{orig} and s_{low} depending on how much the intensity level of the higher frequencies should be attenuated in the resulting combined signal.

To incorporate the earlier described monaural cue, the signal is filtered depending on the relative azimuth of a source with respect to the listener's position and orientation. In case a source is in front of the listener, no filtering is applied. When it enters the back hemisphere the attenuation of the high frequencies becomes stronger until it reaches its maximum in case the source is in the absolute back of the listener. In figure 5.4 the intensity level decrease of the higher frequencies Dec_{back} is depicted in dependency of the relative azimuth of the source.

Implementation of monaural cue

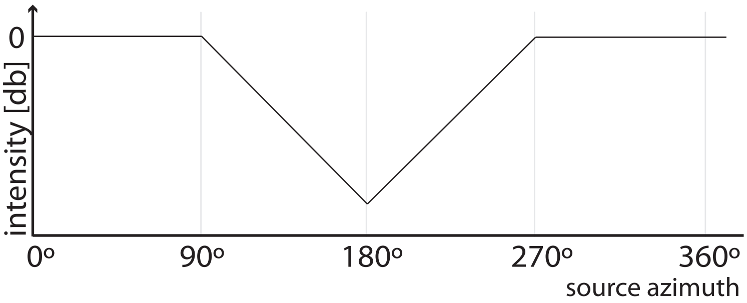


Figure 5.4: Attenuation of the higher frequencies on both ears in dependency to source azimuth.

5.1.3 Method C - Frequency dependent ILD and monaural simulation of the pinnae shadow

As already mentioned, the implementation of OpenAL on the iPhone uses simple stereo panning to create directional cues. This means that the same ILD is used for all frequencies of the signal. Since this is rather unnatural and can lead into diverse problems (see 4.2.3—"Problems of the Spherical Head Model"), we want to test a more advanced auralization method which uses different ILDs for different ranges of frequencies and thereby is closer to the real phenomenon. In Loomis et al. [1991] the authors describe a similar approach but they have in addition incorporated frequency independent ITDs. Since our method lacks the last binaural cue due to reasons described later on, we have to evaluate whether the proposed method will be an improvement in relation to the previously described rendering method.

Different ILDs for different ranges of frequencies

If we review the ILD-to-frequency function of the spherical head model, we notice that in case of all directions the function can roughly be described by the following: The ILD is close to zero for low frequencies; around a frequency of one kHz the ILD rises with a high slope up to nearly the maximum ILD; after this fast ILD increase, the ILD fluctuates around the maximum ILD with only a slight overall growth for rising frequency (see 4.2).

To approximate this behavior we incorporate the lowpass filter described in the previous section again (5.1.2). But now we apply different attenuations of the high frequencies for the left and right ear (Dec_{left}, Dec_{right}) as depicted in figure 5.5 and described in the following:

When the source is located in front of the listener she hears the unfiltered signal on both ears, therefore $Dec_{left} = Dec_{right} = 0$. When the source moves into the left hemisphere of the listener, the attenuation of the high frequencies on the right ear Dec_{right} increases linearly until the high frequencies are attenuated with a maximal value at azimuth = 90^o. Thereby Dec_{left} stays zero. Moving into the back of the listener, the attenuation of the high frequencies decreases linearly again. The right hemisphere is handled analogically. Additionally, the monaural cue for sources in the back of the listener which is described in the previous section is incorporated. Therefore Dec_{back} is added to Dec_{left} and Dec_{right}.

Different attenuations of high frequencies for left and right ear

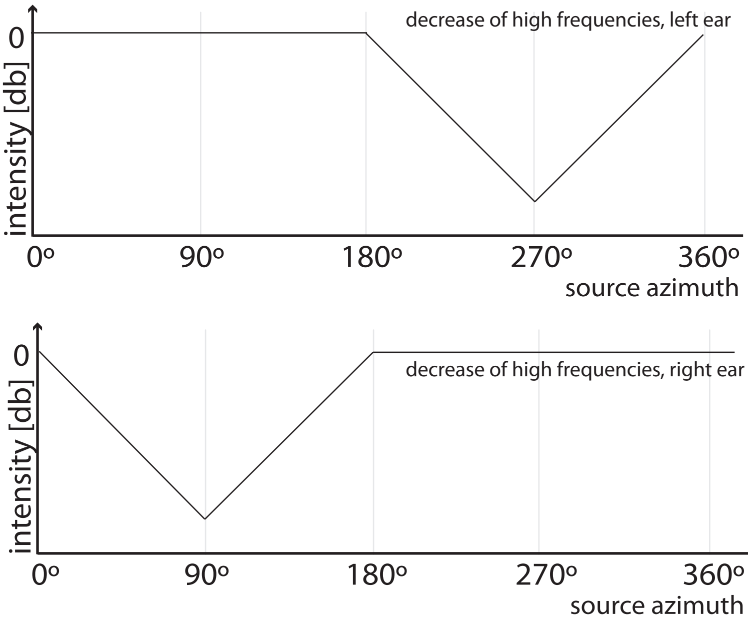


Figure 5.5: Upper picture: Attenuation of the high frequencies on the left ear in dependency of source to listener azimuth. Lower picture: Attenuation of the high frequencies on the right ear in dependency of source to listener azimuth.

We also planned to incorporate ITDs as proposed by the spherical head model (4.6). But we recognized that when shifting the current time-progress of an audio clip in OpenAL with a high frequency (10Hz), unpleasant audio artifacts are created and we had to remove the ITDs again.

ITDs not possible with OpenAL

5.1.4 Reverberation and Volume Distance Attenuation

One of the reasons why the listening experience with our first prototype felt unnaturally was that no room acoustic effects were perceivable. To reduce this problem we incorporate reverberations into the auralization. Besides increasing the perceived realism of the scenario, it is a second aim to provide the listener with an additional cue which enables her to better estimate the distance of a source. As described in 4.4.2—"Perceived Distance", the ratio between reverb and direct signal serves as an absolute distance cue.

Increase realism

Absolute distance cue

Due to the reason of missing filtering possibilities as mentioned several times, we are not able to create the reverb dynamically at runtime. Therefore, we use the approach of using preprocessed audio clips again. For every audio clip which is to be auralized we use a second clip which contains the original clip shaped with a reverb filter and is played to the original clip synchronously. To serve as absolute distance cue, the reverb audio clip is auralized with a volume independent from distance whereas the volume of the original clip is derived with the Inverse Distance Clamped Model (3.1.1—"OpenAL") as depicted in figure 5.6. Additionally, both audio clips were faded out completely for distances at which the original clip is masked by the signals of the other sources in the scene. This distance depends on the setup of the sources in a scene. In case of the small scenarios used in our tests it was approximately 12 meters.

Preprocessed reverb audio clips

Additional fade-outs after masking

The filtering is done offline with an external software and the used reverb filter is adjusted to simulate a room environment close to the place where the target application will be located, in our case the Coronation Hall in Aachen, Germany.

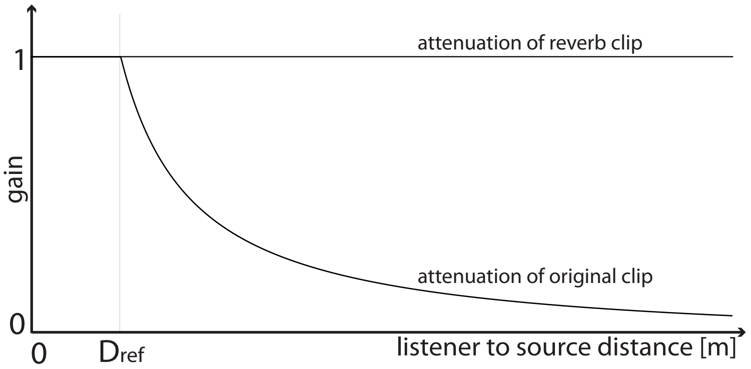


Figure 5.6: Attenuation of the original audio clip and the reverb audio clip.

5.2 Preliminary User Study

During the implementation we ran a preliminary user study to get a fast feedback about the implemented auralization methods. With a navigation performance test we intended to find out if the proposed spatialization methods are an improvement on the method evaluated in the initial test with respect to the concept of navigating solely by auditive cues. We also wanted to test, whether the incorporation of reverberation leads to a more natural listening experience and a better perception of the relative source distance. Additionally, we tested the idea of experiencing a virtual coronation feast by walking through a virtual audio landscape.

5.2.1 Navigation Performance Test

Test concept of navigation-by-ear

In the navigation performance test we wanted to check how well the proposed auralization methods enable the user to navigate solely by auditive cues. Since this quality is not directly measurable, we measured the time a subject needed in a navigation task and used this as measure for the quality of the different methods.

Methodology

The task which the user had to complete is a homing task as proposed by [Loomis et al. [1991]]. It is described in the following: A virtual sound source was presented to the subject at a specific distance and in a random direction to her. The subject then was advised to navigate to the source position unknown to her solely by the spatial cues. When she thinks that she is in a radius of about one meter around the source, she was instructed to give a verbal signal. The starting distance thereby was always eight meters and the source emitted a continuous speech signal.
This procedure was done several times for one auralization method and then repeated for the other methods in a pseudo randomized order.

Subjects navigate to virtual sound sources

To be able to compare the navigation performance of the subjects with the different auralization methods we recorded the time the subjects needed to find a source and the accuracy of determining its position during the tests. Additionally, we wanted to gain some qualitative information in conversations held after and during the tests. The interview after the test was semi-structured by a preselected set of questions and topics. The discussed topics were for example, what was felt to make the navigation easier and what more difficult; whether a delay between motion and audible feedback was noticed; or if the headphone respectively the cables were disturbing.

Measurement of completion time and accuracy

Semi-structured interview

To track the subject's position and orientation we could not use the Vicon system anymore, since the spatial dimensions of scenarios we wanted to evaluate exceeded the trackable area of the available Vicon system. Luckily, the hardware-part of the final system which is responsible for determining the listener's head orientation, was available to that date. An electronic compass module [Heller [2008]] mounted on the top of a headphone, sending the orientation of the head with a rate of 10 Hz by cable to the software prototype, was incorporated.
To keep track of the subject's position we applied a Wizard of Oz approach and used the spatial view/control of the prototype implementation to continuously update the listener's position by hand. To avoid that the project mem-

Electronic compass module

Wizard of Oz position tracking

bers in position of the wizard were influencing the results, they did not get to know or see the sources' positions of the current evaluated scenario.

Results

All participants were able to complete the navigation task under all conditions. We observed a strong tendency that the subjects performed better with our advanced auralization methods than with the basic OpenAL rendering, which we tested in the first prototype. Nevertheless, due to the problems described in the following we are not able to say with evidence which method performed best for the navigation task.

Tendency: Advanced methods better than basic OpenAL

After a few subjects we noticed that the Wizard of Oz tracking technique did not perform accurately enough to be sure that it does not influence the evaluation results. The response time to subject movements was to high and led into a refresh rate of the audio rendering which was sometimes too low for the navigation behavior of the listener. The subjects perceived sometimes a jumping listening experience. Some subjects stated that especially in case of large source distances the changes in the perceived source volume induced by their movements were too small. They argued that they were sometimes not able to estimate wether a step in one direction brought them nearer to the source or not. One subject mentioned the missing of cues like, "heiß oder kalt" referring to a children's game called "Topfschlagen" (hit the pot – in this game a blindfolded child has to find a pot with sweets by beating on the floor with a cooking spoon while the others help her by shouting hints like warm or cold). Due to the problems with the manual tracking, we are not sure wether this is caused by an unresponsive position feedback or by a too low slope of the attenuation curve. Since the navigation problems mentioned by the subjects could have a strong influence on the measurements, we decided to abort the study after five participants.

Wizard of Oz not accurately enough

Missing navigation cues

A second problem we encountered with the test setup, was that most subjects tried to find the exact position of the sources before they signaled that they reached the source. After reaching a radius of approximately one meter around the source, many subjects began to make very small head movements or leaned back and forwards to locate the source position more accurately. Therefore most subjects spent a large amount of the task time only for localizing the exact source position in the direct proximity of the source.

Most time was spent on localizing exact source position

When asked to decide which synthesis method led to the best listening experience, all subjects preferred the more advanced synthesis methods to the one which we already evaluated in the first prototype. They stated that the latter felt unnaturally due to the complete absence of any reverberation. Nevertheless, it was also mentioned that in case of the other methods the reverb was perceived to be too strong and sometimes hindering the navigation.

More advanced synthesis methods preferred

Most subjects were not able to select a clear favorite from the more advanced methods. Some mentioned that they perceived the lowpassed signals for sources behind the head as being "under water" or "in an airplane, with too high pressure on the ears". Others said, that they were not conscious of the effect. One told us that she used the abrupt change of the sound from normal to "muffled" to determine when she just passed a close source.

Lowpass filtering too strong

5.2.2 Coronation Feast Scenario Test

The coronation feast scenario test was conducted to investigate how our idea of letting the listener experience a virtual coronation feast by walking through a virtual audio space was accepted. To get an impression of how people interact with a virtual audio scenario only, we presented them a small coronation feast scenario and let them freely walk around in it. We wanted to investigate if the presented scenario lead into an immersive experience like history becoming alive. Besides that, the test should reveal where possible problems are, especially with respect to the installation of the system as an interactive exhibit on an historic site.

Test of immersion

Uncover unthought usability problems

Methodology

Free exploration of the scenario

The participants put on the headphones and were instructed to walk around freely and explore the presented coronation feast scenario. The scenario consisted of a source emitting an oration in medieval speak and a source emitting a minnesong. Additionally the user perceived a background feast-atmosphere, since we added a source which ubiquitously emitted a binaural recording from a medieval fair and two sources playing recordings of bar-noise. Latter ones were located in opposite scenario corners with low rolloff factor adding more dynamic into the background ambience.

Selectable auralization methods

While walking around, the subject carried along a controller allowing her to switch between two auralization methods she preferred in the navigation task.

Semi-structured interview

Afterwards, we performed a semi structured interview covering the following topics: naturalness of the experience; immersion into the presented scenario; effortlessness of following the content of the source when being near to it; and wether the subject could imagine the presented scenario as an exhibit in a historic site. Some other topics and questions emerged during the interviews due to the open interview style.

Results

Interesting and vivid information conveyance

The overall feedback regarding the presented feast scenario and the interactive experience was generally positive. All subjects valued it as an interesting and vivid way to submit information about past events at a historic site.

Feeling of changing between different scenarios

Talking about the feeling of being present at a virtual coronation feast, one subject mentioned that he felt like changing between two different scenarios. While being near to the minnesong he felt like moving on a market place with a band playing on a stage and while attending to the oration he imagined to be in a hall standing in the last row of a crowd. Another subject had the similar feeling of walking from one environment to another. He said the reason for this was, that in reality a band or a loud speech would be heard much farer. Nevertheless, he valued this as unnatural but not negative.

Problems in near-fields of sources

During the discussions about the listening experience of the subjects in case of being near a source and paying attention to its content, three subjects mentioned that the reverb was perceived as being too loud and therefore "disturbing". Some subjects stated that the continuous changes in the audio signal due to the head movements and spatial rendering were perceived as annoying and confusing. With regard to the usage of the system in a historic site, the wish of being able to look around without being distracted by strong changes in the signal, was mentioned. One subject said about this point: "One does not want to stare at the same location the whole time". Some subjects mentioned that it became difficult to understand and follow the content of the oration, when the vocal part of the minnesong started and originated this on the interference of the two speech signals.

5.3 Discussion and Improvements

Too low slope of distance attenuation curve

Some subjects mentioned that they sometimes were not able to perceive a change in the volume of an approached source in case it was still far away. Although we were not sure whether this was caused by problems with the Wizard of Oz tracking, we reexamined the distance attenuation curve of the source volumes. We observed that when being at a source distance of eight meters and making a step towards the source, the induced volume change is below the minimal difference in loudness which is perceivable by a human. Since the participants mentioned that a rising volume was an important hint for them that they were going into the right direction, we decided to reconfigure the distance attenuation curve to have a higher slope.

Additional linear decrease

We therefore informally evaluated different slopes in internal tests, e.g., a 9dB decrease per distance doubling; an additional linear increase; a complete linear curve and combinations of those. We got the best results with an additional linear decrease of one dB per meter to the natural attenuation. In case of the methods which used the lowpass filter, we utilized the filter to incorporate the additional linear changes. Therefore, only the frequencies above 1000 Hz are attenuated with rising distance. This means, that we come closer to the behavior of signals in a room environment as described in (4.4.1—"Perceived Signal") and incorporate the spectral distance cue as explained in (4.4.2—"Perceived Distance").

Reduction of reverb and monaural pinnae shadow

By taking the other results into consideration, we further reexamined our auralization approaches. After several internal informal tests, we decided to reduce the overall volume of the reverb and to reduce the maximal strength of the monaural pinnae shadow. We think that the latter will not decrease the navigation performance but lead to a more realistic listening experience.

Additional auralization method

As already mentioned, some subjects stated that they experienced the lowpass filter as unrealistic. Although we had already reduced the strength of the filtering, we decided to implement another auralization method. This method, called Method D later on, uses the simple OpenAL rendering to create the directional cues, but incorporates the same reverb and spectral cues as the other more advanced methods. Thereby, we wanted to be able to check how much the incorporation of the lowpass for sources in the back of the listener improves the navigation and whether this increases the perceived realism of the listening experience.

Reduction of directional cues in near-field

During the coronation scenario test, several subjects mentioned that when being close to a source and listening to its content the strong changes in the audio signal caused by the spatial cues disturbed their listening experience. Therefore, we think that it could be an advantage to reduce the directional cues incorporated into a signal of nearby sources. But it has to be explored whether this increases intelligibility and listening comfort, and whether this is not perceived as unnatural or irritating.

Ensure intelligibility in near-field

A second thing that was mentioned by the subjects was, that when being near to one of the sources and listening to its content, the other more distant sources were still perceivable with a volume that reduced the intelligibility of the currently heard content. Therefore, we think that it could be an advantage to ensure, that when a listener is in the near-field of a source, no other more distant source is louder than a specific threshold value.

One possibility to obtain this would be, to adapt the slope of the distance attenuation curves such that the volume of each source is below this threshold at the near-field of each of the other sources. But we have to take into consideration that in the planned target scenario the sources are not evenly distributed. This would mean, that we either use a different attenuation curve for each source or that we use the attenuation curve which is adapted to the two sources which have the smallest distance in the scenario for all sources. In the first case the distance perception would be inconsistent, e.g., two sources which are perceived with the same volume have different distances to the listener. In the second case large gaps would exist between more distant sources.

Different design considerations

Therefore we wanted to explore a different way. If we look at the distance attenuation curve used until now, we notice that on the last few steps towards a source the curve has a high slope. The idea is now to fade out the more distant sources on this last few steps to the earlier mentioned threshold of easy intelligibility. Our thought behind this was, that at this distance the listener's locus of attention is on the nearly reached source and that the faster than normal decrease of the more distant sources is masked by the rapid volume increase of the approached source. But we have to evaluate whether listeners will perceive this as unnatural or confusing.

Fade-out of distant sources in the near-field

Chapter 6

Final User Study

In this section we will present the final test of our auralization approaches, and we will describe the evaluation of the ideas for the improvement of the listening experience in the near field of a source. We will describe the used methodologies and the obtained results. All tests were conducted with a UBIsense position tracking system [UBIsense[1]].

6.1 Test of the Auralization Methods

Two levels of evaluation

With the final test we wanted to check the qualities of our proposed auralization methods with respect to perceived realism and how well the methods enable the listener to navigate in a virtual audio scenario. As proposed by Stanney [2002], we evaluated the methods on two levels. The primary level focuses on measuring the performance of the subjects in a navigation task. In the secondary level we used a questionnaire to obtain information which should support us in interpreting and elaborating the level of performance.

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

6.1.1 Navigation Performance Test

In the navigation performance test we wanted to check how well the proposed auralization methods enable the user to navigate solely by ear in a scenario with multiple competing sound sources. Since this quality is not directly measurable, we measured the time a subject needed in a navigation task and used this as measure for the quality of the different methods. An important aspect of the test was to evaluate the navigation time in a scenario which is close to the target scenario.

Methodology

Scenario with multiple competing sound sources

Due to the problems in the earlier described pilot navigation performance test we modified several aspects of the test design. In this test we presented a scenario with multiple sound sources to the subjects (see figure 6.1).

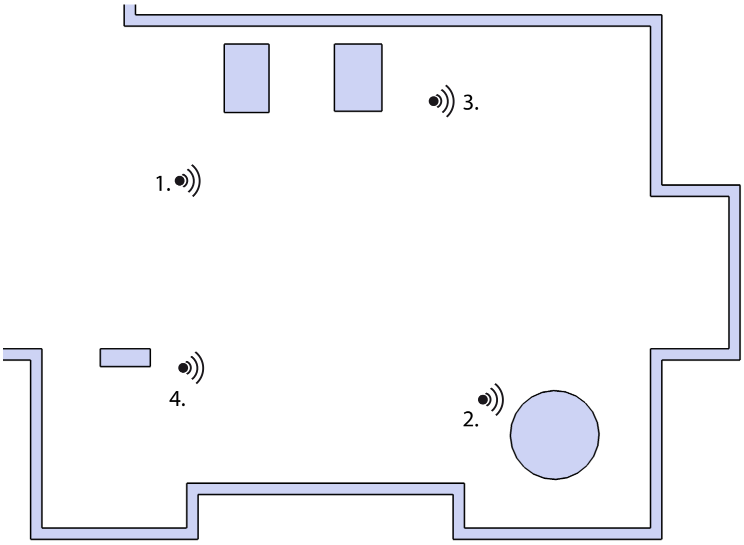


Figure 6.1: Source setup of a navigation task scenario in the ground plan of the test environment. The blue areas are obstacles or walls. The numbers specify the order in which the sources have to be approached.

The sources played several audio clips with speech recordings simultaneously. For every source the recording of a different speaker was used. The content of the sources were distinguishable from each other by the spoken content. One source, for example, only played first names, another one only random numbers. Before each navigation task we gave a note with a list of numbered terms to the subject. Each term described with one word the content of one of the sources in the scenario, e.g., first names. We then asked the participant to navigate to the sources in the given order. Afterwards we played a "bling"-sound to her and told her that when she reaches the current target source she will hear this sound and shall navigate to the next source on the list. The "bling"-sound was triggered by a subject-to-source-distance lower than one meter.

Navigation to sources in given order

Every subject performed a navigation test with all four auralization methods.

- Method A - Head related transfer function
- Method B - Frequency independent ILD and monaural simulation of the pinnae shadow
- Method C - Frequency dependent ILD and monaural simulation of the pinnae shadow
- Method D - Frequency independent ILD

For every method a different scenario, i.e., with different source positions but the same contents and speakers, was presented. We also changed the order in which the sources had to be approached with each method.

The scenarios contained four target sources arranged on a 13x10 meter large area. We placed additional sources around the target sources. Initially we designed the scenarios twice as large, with 7 target sources. But as we tested the tracking system on the test site, it showed up that the tracking range was smaller than expected, thus we had to decrease the size of the scenario.

Smaller scenarios than planned

In every navigation task we measured the time from the moment we presented the scenario to the subject until she reached the last target source. After the last navigation task we asked the subjects several questions.

Users

A number of 10 voluntary users, seven are male and three are female, participated in the test. The ages differed from 23 to 60 and all affirmed that they had full hearing abilities and no experience with audio augmented reality.

Due to the earlier described differences in the auralization methods we formulated the following hypothesis:

- H1: Subjects will need less time with Method A than with Method D.
- H2: Subjects will need less time with Method B than with Method D.
- H3: Subjects will need less time with Method C than with Method D.

Results

The resulting task completion times range for Method A from 45.3 seconds to 204.3 seconds with a mean of 80 seconds and a standard deviation of 50 seconds. In case of Method B we measured times between 53 and 111 seconds with a mean of 69.6 seconds and a standard deviation of 25 seconds. Method C produced times from 44 seconds up to 175 seconds, with a mean of 99.2 seconds and a standard deviation of 44.9 seconds. In case of Method D we measured times in the range of 64.5 to 199.4 seconds. The mean for this times is 88 seconds and the standard deviation is 43.6 seconds (see table 6.1).

Comparison of methods

To compare the different methods with each other we calculated the time differences between each two methods for every person. The mean of the personal time differences for each two methods is visible in table 6.2, together with the standard deviation and the p-values.

Auralization Method	Mean Time	Std. Dev.
Method A	80.0s	50.0s
Method B	69.6s	25.0s
Method C	99.2s	44.9s
Method D	88.0s	43.6s

Table 6.1: Mean navigation task times and standard deviation.

Compared Auralization Methods	Mean	Std. Dev.	p-value
Method A time - Method D time	-8.0s	19.3s	0.123
Method B time - Method D time	-18.3s	20.8s	0.055
Method C time - Method D time	11.1s	32.3s	0.165
Method A time - Method B time	10.3s	31.7s	0.179
Method B time - Method C time	-29.5s	39.1s	0.026
Method A time - Method C time	-19.2s	42.8s	0.107

Table 6.2: Means of the differences between the navigation task times of different methods. Paired t-test, n=9.

Discussion

The mean differences in the subjects' performances between the methods were in most cases not significant. We think that one reason for this is the small size of the used scenarios. We think that in case of using larger scenarios with more target sources and testing in a larger environment would increase the significance of the results.
Another reason is probably that only nine time measurements per auralization method are measured (the measurements of one participant are not usable since problems with the tracking system occurred while he performed the navigation task and we are not sure whether it did influence his navigation performance). Furthermore we noticed a slight tendency that some of the used scenarios lead to higher times (e.g. difference $\bar{x}$=13.6s, p=0.12). This lets us suggest that it has a higher difficulty with respect to the navigation task and therefore caused a noise increase in the measures.

Tendency: Method A and B better than C

Nevertheless, the results reveal the tendency that subjects perform better with Method A and Method B than with Method D. In case of Method B the average time difference to Method D was 18.3 seconds (σ=20.8s) but it was only weakly significant.

Tendency: Method C worse than D

In contrast to our hypothesis, the results show a tendency that the performance with Method C was even worse than with Method D. Comparing Method C with Method B, the first even lead to significantly higher times. A possible explanation could be that Method C incorporates smaller ILDs than Method B and Method D. These are more realistic but lead to smaller volume changes in the perceived audio signals on the left and right ear when turning the head. Since many participants stated (see 6.1.3—"Qualitative Results") that they generally used the changing in the perceived volume to locate sound sources, this may be a reason that Method D lead to longer task completion times.

Some subjects generally faster than others

During the test we observed that some subjects generally performed faster than the others. To check wether this holds true, we calculated the personal mean time with respect to the four methods for every person. We noticed that the personal average times vary largely between the subjects. They ranged from 54.8 seconds to 172.6 seconds. The personal standard deviation of each person was in average (20.3 seconds) rather small. Therefore, we think that some participants were generally more skilled than others with respect to the navigation task. In consequence they were proportionally faster under all conditions.

Normalize completion times

To reduce the influence of different navigation abilities we normalize the measured values according to the personal skills. To achieve this we divide the measured times of each subject by her personal average time. By this we are able to estimate the factor of how much a method increases the navigation performance with respect to the average performance of the person. In table 6.3 we see the average and standard deviation of the normalized times for each method. To check whether this mean values differ significantly between the methods, we also perform a t-test for every method pair. The result is depicted in table 6.4.

Significant: Method B better than D

The normalized task times reflect the already mentioned tendencies. The mean difference of the normalized task times needed with Method B and Method D is 0.19 (σ=0.23, p=0.017) and significant. The mean difference between Method A and Method D is 0.12 (σ=0.24, p=0.08) and weak significant.

Auralization Method	Mean	Std. Dev.
Method A normalized time	0.92	0.21
Method B normalized time	0.85	0.16
Method C normalized time	1.18	0.31
Method D normalized time	1.04	0.13

Table 6.3: Mean normalized navigation task times and standard deviation.

Compared Auralization Methods	Mean	Std. Dev.	p-value
Method A - Method D	-0.12	0.24	0.080
Method B - Method D	-0.19	0.23	0.017
Method C - Method D	0.14	0.37	0.147
Method A - Method B	0.07	0.21	0.180
Method B - Method C	-0.33	0.43	0.025
Method A - Method C	-0.26	0.50	0.077

Table 6.4: Means of the differences between the normalized navigation task times of different methods. Paired t-test, n=9.

6.1.2 Questionnaire

To gain more specific information about the performance of the different auralization methods, we conducted a questionnaire. The answers should help us to better understand the results of the navigation performance task and to find out which qualities improve the navigation performance. Furthermore, we wanted to gain information about how realistic the listening experience with the different methods was perceived by the subjects. The questionnaire should provide us with more differentiated informations and arguments to discuss which auralization method should be used in the target application. Finally, we wanted to gain information for a further improvement of the proposed auralization methods.

The participants were asked to rank the auralization methods with respect to the statements S1-S6 listed in the following.

- S1: "The navigation between the speakers is easy."

For this statement we had the same hypothesis like for the navigation performance test. We expected that Method A, Method B, Method C will become better ratings than Method D.

- S2: "The decision whether a source is in front of or behind me is easy."

Since Method B and Method C artificially enhance the perceivable cues for front-back-differentiation we formed the hypothesis for this statement that subjects will rate Method B and Method C higher than Method D. Because Method A uses non-individual HRTFs, the problem of front-back-confusion is likely to occure (see 4.2.4—"Anatomically Correct Transfer Function"). Therefore we thought that Method B and Method C will also perform better than Method C.

- S3: "I always have the feeling that the sources are very near or inside my head."

The externalization of sound sources depends on how similar the spatial cues in the presented signal are to the cues which would occur with a real sound source (see 4.2.3—"Problems of the Spherical Head Model"). Therefore it would be likely that Method A will perform better than Method D, but since Method A uses non-undividual cues we are not confident about this hypothesis. Method C extends Method B and Method D by incorporating frequency dependent ILDs, and thereby gets closer to reality. Therefore we form the hypothesis that Method C will become better ratings than Method B and Method D.

- S4: "I have a clear spatial conception of the source locations."

For this statement we again formed our standard hypothesis that Method A, Method B, and Method C will be rated higher than Method D. Furthermore, we are interested whether a clear spatial concept is essential for a good navigation performance. Therefore, we will look for a correlation between this statement and the ratings respectively to the navigation.

- S5: "When I am hearing multiple speakers concurrently, I can segregate and understand them without effort."

As described in 5.1.2—"Method B - Frequency independent ILD and monaural simulation of the pinnae shadow" we expect that spectral shaping of some of the concurrently heard speech signals enhances the ability to segregate the signals from each other and increases the intelligibility of the content. Therefore we expected that Method B and Method C will be better rated than Method D. Nevertheless, true binaural unmasking of speech signals (5.1.2—"Method B - Frequency independent ILD and monaural simulation of the pinnae shadow") originates from differences between the ITDs of the signals. Since only Method A incorporates ITDs we thought that it would be better rated than Method B, Method C, and Method D.

- S6: "The listening experience gives me the feeling of walking among real speakers"

Since the lowpass filtering of background sources is quit artificially we would assume that Method B and C become lower ratings than the other two methods.

Methodology

Rating of methods by scale from one to ten

The following test was performed directly after the navigation performance test. We presented to each participant a statement, e.g., "The decision whether a source is in front or behind me is easy.". Additionally, we presented a sheet of paper to the subjects with a scale ranging from one to ten. One corresponding to "I strongly agree." and ten corresponding to "I strongly disagree.". Four small physical tags were laid on the sheet, each one associated with one auralization method (see figure 6.2). Furthermore, we gave an iPod Touch displaying four buttons to the subjects. Each button was associated with one auralization method and activated the respective method when it was touched. The buttons were designed to match the design of the physical tags. We subsequently presented a scenario with several sound sources to the subjects. Afterwards, they were advised to walk around freely in the scenario, switch between the different methods, and step-by-step move the physical tags until they correspond to the their opinions to the presented statements. We always shortly explained the meaning of the presented statement to the subjects and asked whether they understand the question and the task. We re-

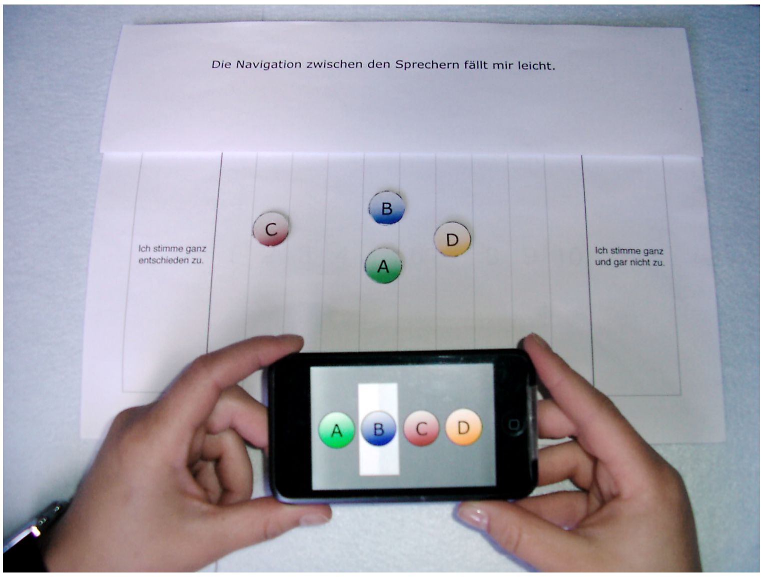

Die Navigation zwischen den Sprechern fällt mir leicht.
Ich stimme ganz entschieden zu.
C
B
D
A
Ich stimme ganz und gar nicht zu.
A
B
C
D

Figure 6.2: Setup of the questionnaire study.

Possibility to freely walk around and switch between methods peated this procedure for all six statements S1 - S6. The association between the auralization methods and the physical tags was changed in a semi randomized order between each subject.

Results

The mean times and standard deviations for each question and method can be found in table 6.5 and 6.6. Additionally the differences between each pair of methods are depicted in table 6.6 and 6.7 together with their standard deviations and p-values. In figure 6.3 we can see the results plotted as box-and-whisker diagram.

S1: "The navigation between the speakers is easy." In case of statement S1, our hypothesis that Method A and Method B are better rated than Method D is supported by the study. Method A was rated significantly better than Method D with a mean difference of 2.4 (σ=2.37, p=0.005). In case of Method B the mean difference was 1.9 (σ=2.42, p=0.0175) and significant.
Although Method A generally obtained good scores ($\bar{x} = 3.5, Q_{0.25} = 4, Q_{0.5} = 3, Q_{0.75} = 2$) it was rated very badly in two cases (8,6). In contrast the ratings of Method C have generally a strong deviation, as the lower quartile is 7 and the upper quartile is 2 (see diagram 6.3). There is no significant difference between the scores of Method C and Method D.

Significant: Method A and B better than C.

S2: "The decision whether a source is in front of or behind me is easy." The hypothesis that Method B and C become a better ranking than Method A and Method D with respect to statement S2 is fully supported by our study. Method B was rated better than Method D with a mean difference of 3.1 (σ=3.38, p=0.0088) and better than Method A with a mean difference of 3 (σ=3.62, p=0.0139). The mean difference between the ratings of Method C and D was 3.5 (σ=4.2, p=0.0135) and 3.4 (σ=3.17, p=0.0039) between Method C and Method A.

Significant: Method B and C better than A and D

Tendency: Method C better than B and D

Weakly significant: Method A better than D

S3: "I always have the feeling that the sources are very near or inside my head." Concerning S3, our hypothesis that Method C will be rated better than Method B and Method D could not significantly be proven by the study. Nevertheless, the results reveal a tendency to support our assumptions (please remind that in case of this statement high values are better than low ones). The median of Method C ($Q_{0.5}$=7) is two points larger then the medians of Method B and Method D. Also the mean of the score differences between the Method C and Method D ($\bar{x}$=1,7), resp. Method C and Method B ($\bar{x}$=1.1), reflect this tendency. However the differences are only weakly or not significant at all. Besides that, Method A scored higher than Method D with a mean difference of 1.4, nevertheless, with a p-value of 0.06 this value is only weakly significant.

Subjects: Problems in distinguishing between methods

S4: "I have a clear spatial conception of the source locations." In case of S4, the scores of all conditions had a large deviation (see diagram 6.3). The mean values of all methods are close to 4 with only a small deviation. Some subjects mentioned during the test that they had problems with this statement. They stated that they did not perceive a difference between the four methods with respect to S4. Two participants therefore rated all four methods identically, one subject ranked all the methods with a score of 8 and another subject with a score of 1.

Weakly significant: Method A better than C and D

S5: "When I am hearing multiple speakers concurrently, I can segregate and understand them without effort." Our hypothesis for S5 was that Method A will be receive better ratings than the other three methods. This is not supported by the study. Nevertheless, the mean differences between Method A and the other three methods show a slight tendency that Method A performs better with respect to S5. Comparing Method A and Method D, the mean of the difference is only weakly significant ($\bar{x}$=1.5,σ=2.84,p=0.0645). The same holds true for the comparison of Method A and Method C ($\bar{x}$=1.5,σ=3.41,p=0.098). In case of Method B the difference was even smaller. Our hypothesis that the two Methods incorporating the lowpass filter, Method B and Method C, are better rated than Method D with respect to this statement could not be supported by this study. A possible reason may be that the shaping of the signals which came from sources in the back of the subjects was perceived to reduce intelligibility of the sources in the back.

S6: "The listening experience gives me the feeling of walking among real speakers" The results for S6 do not show any significant differences between the Methods. Many participants mentioned they had difficulties to perceive any distinction between the methods with respect to this statement. Half of the subjects each gave nearly the same score to all Methods. Albeit one subject stated that the listening experience was quite unrealistic with all methods since the reverb sounded more like being in a church or a large hall. She therefore rated all methods with a 7. Other subjects also mentioned that the reverb is not appropriate for the room size.

Differences between methods not significant

6.1.3 Qualitative Results

Most participants (8/10) stated that they needed only a few steps, i.e., until one source became louder, to become familiar with navigating solely by the spatial cues. One participant stated that this familiarization time lasted until the second sound source was approached and he explained, that he was confused by hearing to many concurrent voices in the beginning. Another subject revealed that he was not sure how his movements induce changes in the perceived audio signal during the hole first navigation task. A possible explanation for this is, that at her first task the Method D was active, which the subject rated very low in the questionnaire. All subjects stated that after these short times of familiarization the interaction was perceived to be natural.

Short familiarization time

Interaction perceived as natural

Some subjects (4/10) mentioned that when starting the experience a short suggestion to move around could be useful since then the heard sound would change and the interaction would become clear. Four subjects mentioned a feeling of confusion in the moment when the experience started for the first time. They stated that this feeling was caused by being confronted with too many voices directly

Suggestion to start moving in beginning

Most subjects felt oriented during the test

at the start. Most subjects (9/10) stated that they felt orientated after a few steps. One subject stated that she sometimes felt disoriented when being in the middle of several sources. Later the same subject reported during the second part of the study, that she always felt oriented, when only three sources were hearable at once. Nevertheless, it was also considered by two subjects, that when using the system over a longer time period, the navigation-by-ear between too many concurrently competing sources could become exhausting.

Infrequent system latency

Two participants perceived infrequently a latency of the spatial sound to head movements, and one judged it to be hindering if it occurred. The same two subjects and another perceived a delay in the update of the sound according to their position when making fast movements. The cable which connected the headphone with the laptop running the prototype was perceived as disturbing by two participants. One mentioned that he sometimes felt hindered by the cable. All participants stated that they did not perceive motion sickness under any condition.

No motion sickness

Five participants found the reverb unrealistic for the room, one stated that it made the navigation more difficult. Only one subject perceived the distance attenuation as slightly too strong to be realistic.

Volume changes used for navigation

Four subjects stated that they used the strong volume attenuation at one of the ears when a source is located, e.g., to the far right of the their heads, to better estimate the direction of the source. Three stated that the changes in the volume of a source were the best cues to estimate the direction of the source and whether one is approaching it or not.

6.1.4 Conclusion

The study revealed that all participants were able to navigate in the presented audio scenario solely by the auditive spatial cues with all the tested auralization methods. All subjects, which made their first experience with the system when Method A, Method B, or Method C were active, needed only a few steps to familiarize with the interaction. After this the interaction with the audio scenario was perceived to be natural.

Only few steps to familiarize with navigation-by-ear

Nevertheless, for the final application we suggest that a short hint is given at the beginning of the interaction to ensure that all visitors make these few steps which are needed to familiarize with the concept of the interaction. Besides that, we think that it would be beneficial that when starting the application only one or two sources should be hearable to allow an easy start into the interaction.

Short hint to move in the beginning

Furthermore, our study revealed that our more advanced auralization methods, Method A and Method B, were preferred by the subjects to the more simple one, Method D.

Method A and B preferred to simple rendering

The results have also shown that the monaural cues for sources in the back of the listener, which are incorporated in Method B, reduce significantly the problem of front-back-confusion without decreasing the perceived realism compared to the basic stereo panning technic incorporated by Method D.

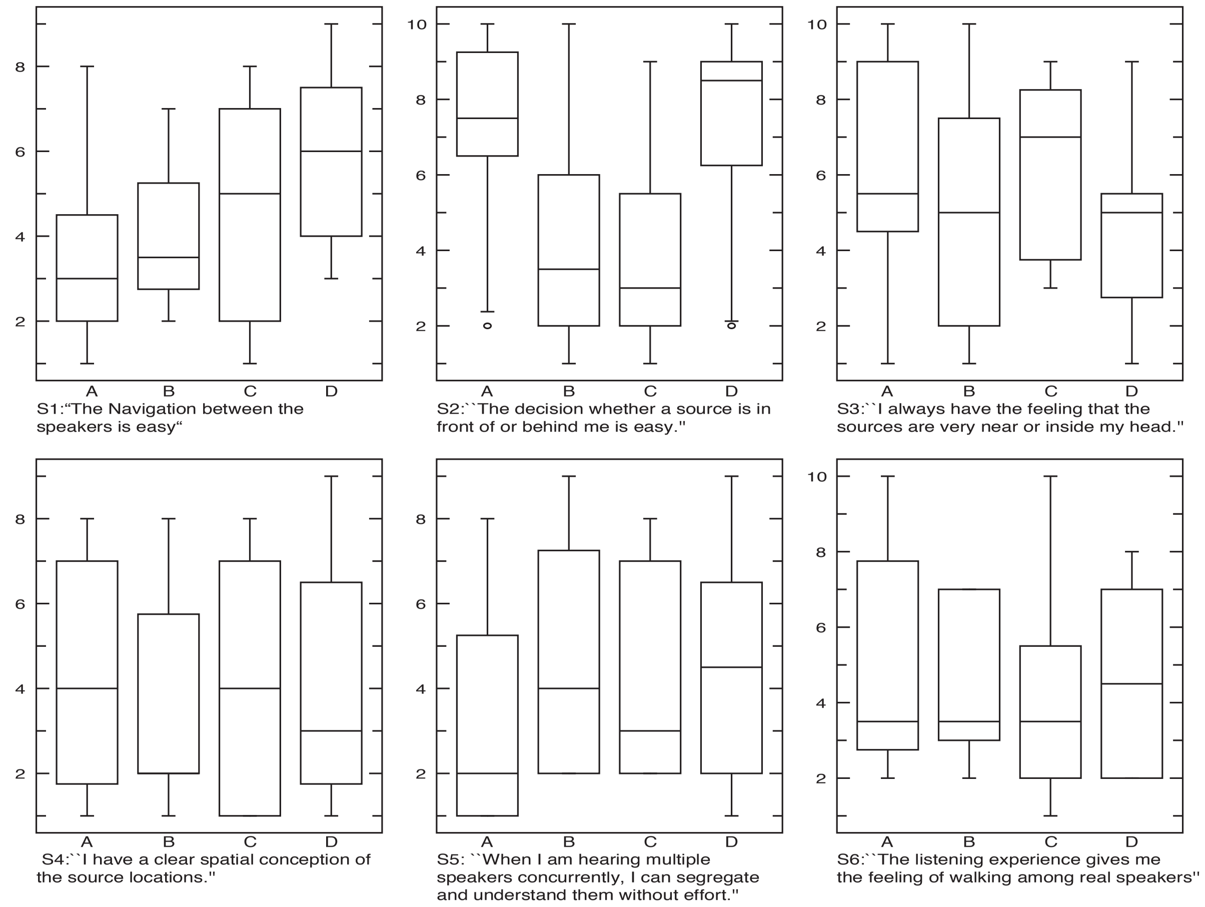
reduced spatial ambiguity without decreased realism

Although Method A performed very well in average, we observed that in case of some few subjects the navigation performance broke down. We think that the reason for this is, that Method A uses non-individual HRTFs in the auralization. As described earlier (see 4.2.4—"Anatomically Correct Transfer Function"), the foundation for providing a listener with accurate spatial cues by an HRTF is, that this HRTF comes close to the individual HRTF of the listener. Although the incorporated HRTF appears to be appropriate for the majority of our subjects, it seems that in case of some few subjects the difference between their personal HRTF and the HRTF incorporated in Method A leads to a navigation performance well below the average.

Furthermore, the study revealed that Method C is no improvement to our simple auralization method. As a matter of fact, the navigation performance test revealed a tendency that Method C performs even worse than Method D.

In conclusion we can say, that the auralization approach of Method B is the most appropriate of all the tested approaches for our project.

Method B most appropriate for our project



S1:“The Navigation between the speakers is easy“

S2:``The decision whether a source is in front of or behind me is easy."

S3:``I always have the feeling that the sources are very near or inside my head."

S4:``I have a clear spatial conception of the source locations."

S5: ``When I am hearing multiple speakers concurrently, I can segregate and understand them without effort."

S6:``The listening experience gives me the feeling of walking among real speakers"

Figure 6.3: The scores given to the auralization methods in the questionnaire plotted as box-and-whisker diagram.

S1: "The navigation between the speakers is easy."

Auralization Method	Mean	Std. Dev.
Method A	3.5	2.21
Method B	4.0	1.70
Method C	4.6	2.63
Method D	5.9	2.02

S2: "The decision whether a source is in front or behind me is easy."

Auralization Method	Mean	Std. Dev.
Method A	7.4	2.46
Method B	4.4	3.00
Method C	4.0	2.49
Method D	7.5	2.55

S3: "I always have the feeling that the sources are very near or inside my head."

Auralization Method	Mean	Std. Dev.
Method A	6.0	2.83
Method B	5.2	2.97
Method C	6.3	2.36
Method D	4.6	2.32

S4: "I have a clear spatial conception of the source locations."

Auralization Method	Mean	Std. Dev.
Method A	4.2	2.66
Method B	3.7	2.63
Method C	4.2	3.01
Method D	4.0	2.79

S5: "When I am hearing multiple speakers concurrently, I can segregate and understand them without effort."

Auralization Method	Mean	Std. Dev.
Method A	3.1	2.42
Method B	4.6	2.59
Method C	4.0	2.40
Method D	4.6	2.67

Table 6.5: Results of the questionnaire I.

S6: “The listening experience gives me the feeling of walking among real speakers”

Auralization Method	Mean	Std. Dev.
Method A	5.0	3.09
Method B	4.5	2.01
Method C	4.2	2.70
Method D	4.6	2.32

S1: “The navigation between the speakers is easy.”

Compared methods	Mean	Std. Dev.	p-value
Method A - Method D	-2.4	2,37	0.005
Method B - Method D	-1.9	2.42	0.0175
Method C - Method D	-1.3	4.20	0.1775
Method A - Method B	-0.5	3.14	0.3
Method B - Method C	-0.6	2.12	0.197
Method A - Method C	-1.1	4.15	0.211

S2: “The decision whether a source is in front or behind me is easy.”

Compared methods	Mean	Std. Dev.	p-value
Method A - Method D	-0.1	2.92	0.458
Method B - Method D	-3.1	3.38	0.0088
Method C - Method D	-3.5	4.20	0.0135
Method A - Method B	3.0	3.62	0.0139
Method B - Method C	0.4	2.55	0.315
Method A - Method C	3.4	3.17	0.0039

S3: “I always have the feeling that the sources are very near or inside my head.”

Compared methods	Mean	Std. Dev.	p-value
Method A - Method D	1.4	2.59	0.06
Method B - Method D	0.6	3.27	0.2881
Method C - Method D	1.7	3.74	0.092
Method A - Method B	0.8	4.42	0.29
Method B - Method C	-1.1	3.11	0.146
Method A - Method C	0.3	3.56	0.3975

Table 6.6: Results of the questionnaire II.

S4: "I have a clear spatial conception of the source locations."

Compared methods	Mean	Std. Dev.	p-value
Method A - Method D	0.2	2.07	0.41
Method B - Method D	-0.3	2.45	0.353
Method C - Method D	0.2	2.94	0.4171
Method A - Method B	0.5	3.57	0.334
Method B - Method C	-0.5	2.59	0.2785
Method A - Method C	0.0	2.67	0.5

S5: "When I am hearing multiple speakers concurrently, I can segregate and understand them without effort."

Compared methods	Mean	Std. Dev.	p-value
Method A - Method D	-1.5	2.84	0.0645
Method B - Method D	0.0	2.58	0.5
Method C - Method D	-0.6	2.37	0.2216
Method A - Method B	-1.5	3.41	0.098
Method B - Method C	0.6	1.78	0.156
Method A - Method C	-0.9	3.54	0.221

S6: "The listening experience gives me the feeling of walking among real speakers"

Compared methods	Mean	Std. Dev.	p-value
Method A - Method D	0.4	1.71	0.239
Method B - Method D	-0.1	1.60	0.42
Method C - Method D	-0.4	2.46	0.309
Method A - Method B	0.5	2.64	0.281
Method B - Method C	0.3	2.06	0.327
Method A - Method C	0.8	2.97	0.208

Table 6.7: Results of the questionnaire III.

6.2 Listening Experience in the Near-Field

The preliminary study revealed some problems in situations when the user was next to a source and wanted to listen to its content. One problem was, that other more distant sources were still that loud that they reduced the intelligibility of the nearby source. To reduce this problem we proposed the idea of reducing the volume of other sources when being in the near-field of a source to a specific volume threshold of effortless listening (see 5.3—"Discussion and Improvements"). Since the additional volume decrease leads to divergency from the distance attenuation of the volume used everywhere else, this may also lead in a listening experience perceived as unnatural. Therefore, one aim of this study is to explore if there is trade-off between intelligibility and a natural listening experience.

Volume threshold of effortless listening

A second thing which was mentioned during the preliminary study is that the strong changes in the perceived signal induced by the spatialization were valued as disturbing when being in the near-field of a source and listening to its content. Therefore, we proposed the idea of reducing the influence of head turnings at the nearby sources to achieve a better listening experience. Since this leads to an inconsistency of the spatial cues in this study we want to explore whether a reduction of the spatial cues leads into a better listening experience and whether a good listening experience can be balanced with perceived realism.

Reduce influence of head turnings

6.2.1 Methodology

For this study we again prepared a scenario with several sound sources. Each sound source emitted a speech signal. Three of the sources D1, D2, and D3 contained dialogs taken from a radio play series, each with the same speakers (see figure 6.4). The reason for this was that we observed during earlier tests, that some speakers were easier to understand than others. Since we wanted to compare the intelligibility of the different sources, we wanted to reduce this influence on the results by using the same speakers for all the sources. Although you never have the same speaker at different positions in a real listening experience.

Three sources with dialogs

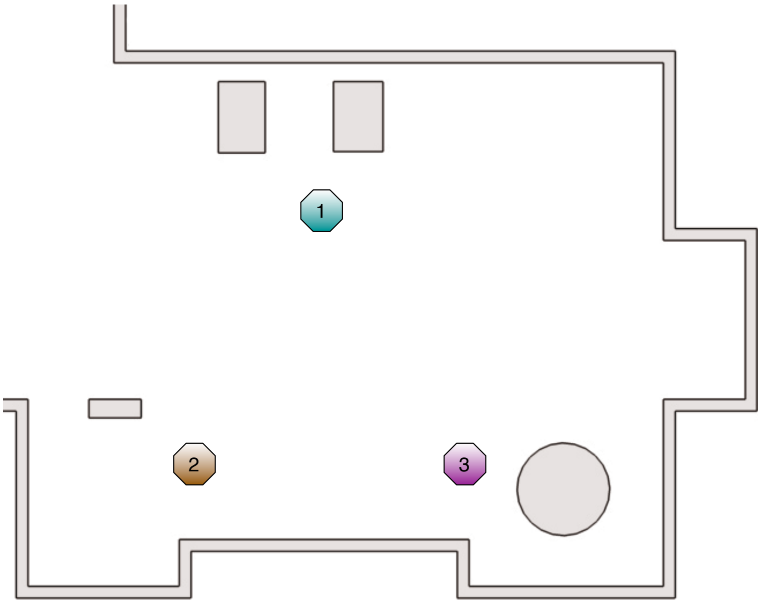


Figure 6.4: Source setup in the floor map plan of the test environment. The grey areas are obstacles or walls. The numbers specify the sources which have to be ranked.

Setup 1: increased volume difference

In a first setup, Setup 1, the sources D1,D2, and D3 differed from each other in loudness with which the other sources were still hearable in the near-field of the specific source. Since the sources had a distance of 8 meters from each other, the generally used distance attenuation curve induced that when being next to one of the sources, D1-D3, the volume difference to the other sources was initially 26 dB. In the near-field of D1, the volume of the other sources was additionally reduced by 6 dB. In the near-field of D3 the volume of the other sources was decreased by additionally 12 dB. In the near-field of D2 there is no additional decrease. In this setup we used Method B to auralize the scenario.

Setup 2: reduced influence of head turnings

In a second setup, Setup 2, the sources D1,D2, and D3 differed from each other in the incorporated directional cues which were used to auralize the source when being close to it. We generally used Method B to auralize the sources and only when the subject came close to one of the sources D1 and D3 we modified the auralization of the approached source. When being near to D1, it was auralized without incorporating the lowpass filtering if the subject turns her head away from the source. When being near D3, it was auralized without any directional cues. The auralization of D2 did not change when coming close to it.

Smooth transitions

The changes to the modified auralizations, which became active when the listener came close to D1 and D2, had a smooth transition. They started to fade in at a distance of 2.5 meters and were fully active in an area with a diameter of 2 meters around the source.

Ground map with sources

Since the participants should rank the different sources with respect to certain qualities we provided a map of the room which showed the sources in a ground map of the test environment and named them with a number (see figure 6.4). Our second idea behind using a map was, that we wanted to keep out the influence of eventual navigation problems or that the wrong sources were approached. Additionally we told the subjects that the listening experience in the direct proximity of each source differs from each other. We did not tell the user in which way since we did not want to bias the subjects.

Subjects rate listening experience on scale

After that we presented one of the two setups to the listener and asked them to rate the listening experience at each source with respect to a given statement on a scale from one to ten. One corresponding to "I strongly agree." and ten corresponding to "I strongly disagree.". The used setup was similar to the setup used in the earlier described questionnaire. The setup only differed in the appearance of the physical tags which were used by the subjects to state their opinion. The appearance now looked like the symbols which were depicting the sources on the map. The following statements were presented:

- S7: "It is effortless to follow the dialog."
- S8: "The listening experience seems to be unnatural/unrealistic."

Subjects move freely between sources

Each statement was presented separately and the participants were recommended to walk around freely between the different sources and to judge each source with respect to the current statement step by step. Then the same procedure was repeated with the other setups and the same statements. To one half of the participants Setup 1 was presented first and to the other half Setup 2 first.

Subjects have already experience

Our subjects were the same as in the earlier described test of the auralization methods. Therefore, they already had some experience with the system and the interaction with a virtual audio space.

6.2.2 Results

Setup 1

In case of the Setup 1 and statement S7, "It is effortless to follow the dialog.", the scores for the source D1 with the volume difference of 32 dB to the volume of the more distant sources ranged from 1 to 4 with a mean of 1.89. The scores for the source D2 with the 26 dB volume difference were distributed from 1 to 9 with a mean of 5.78 (see table 6.8). The ratings for the source D3 with the volume difference of 38 dB ranged from 1 to 3 with a mean of 1.44. Therefore D1 and D3 were rated better than D2. Comparing D1 with D2 the mean of the differences is 3.89 (σ=2.32, p=0.0005) and comparing D2 with D3 the mean of the differences is 4.33 (σ=2.18, p=0.00017), thus the difference is in both cases significant. The difference between D1 and D3 was 0.44 (σ=1.74, p=0.232) and not significant.

Significant: D1 and D3 better than D2

In case of Setup 1 and statement S8, "The listening experience seems to be unnatural/unrealistic.", the participants rated D1 in average with a 6 (σ=3.61), D2 with a 6.33 (σ=3), and D3 with 4.22 (σ=2.05). The mean of the differences between the ratings of D1 and D2 is, with a value of 0.33, rather small (σ=1.58, p=0.272) and showed not to be significant. Comparing D1 and D2 to D3, the first two were rated to be more natural then D3. Comparing the scores of D2 to D3, the mean of the differences is (σ=2.11, p=0.03) and significant. Comparing D3 to D1 it is 1.78 (σ=3.35, p=0.075) and weakly significant.

Significant: D2 better D3

Weakly significant: D1 better D3

Setup 2

In case of both statements the results of Setup 2 do not show significant differences between the sources D1, D2, and D3 (see table 6.8). With respect to statement S7, "It

No significant differences

is effortless to follow the dialog.", the mean score values of all three sources only have small deviations from each other. The mean score for D1 was 3.7 (σ=2.95), the mean for D2 was 3.9 (σ=2.38), and the mean for D3 was 3.5 (σ=1.78). The differences showed not to be significant. In case of statement S8, "The listening experience seems to be unnatural/unrealistic.", the mean scores are slightly more distributed, in case of D1 the mean was 6.6 (σ=3.13), in case of D2 the mean was 5.9 (σ=3.28), and in case of D3 the mean score was 4.7 (σ=2.71). Again the differences showed not to be significant.

6.2.3 Discussion and Conclusion

No enhanced listening comfort

The study was not able to show, that a reduction of the variations in the audio signal which are caused by changes in the orientation of the listener's head lead to a significantly enhanced effortlessness of following the content.

During the test we observed, that the most participants did not move or turn their head while they were listening to the content. Since a fixed head orientation leads into a constant audio signal in case of all sources, D1, D2, and D3, a possible explanation of the result is that the subjects were in most cases not able to perceive any differences between the different sources with respect to an effortless listening.

Drawback of the test setup

In conclusion we may say that the reduction of variations in the signal is not needed. But we are not sure if the mentioned behavior of the subjects was caused by the artificial context of the test and would be different in case of the target application. Since the coronation hall is a much more interesting environment than the environment in which the test was conducted, it is more likely that the listeners will turn their head around to look, e.g., the architecture or artifacts while listening to the content of a source.

Another reason which may cause a different behavior in the target application may be the amount of time the listeners spend continuously at one source. In the test we observed that the participants stayed only short time (approximately 20 seconds) at one source and then walked over to another source to compare them with each other. It is possible that, if the visitors stay longer at one source to listen to its whole content, they may start to look around.

Setup 1:

S7: “It is effortless to follow the dialog.”

Source	Mean	Std. Dev.	
D1	1.87	1.17	
D2	5.78	2.39	
D3	1.44	0.88	
Compared sources	Mean	Std. Dev.	p-value
D1 - D2	-3.89	2.32	0.0005
D2 - D3	4.33	2.18	0.00017
D1 - D3	0.44	1.74	0.23

S8: “The listening experience seems to be unnatural.”

Source	Mean	Std. Dev.	
D1	6.00	3.61	
D2	6.33	3.00	
D3	4.22	2.05	
Compared sources	Mean	Std. Dev.	p-value
D1 - D2	-0.33	1.58	0.272
D2 - D3	2.11	2.89	0.030
D1 - D3	1.78	3.35	0.075

Setup 2:

S7: “It is effortless to follow the dialog.”

Source	Mean	Std. Dev.	
D1	3.70	2.95	
D2	3.90	2.38	
D3	3.50	1.78	
Compared sources	Mean	Std. Dev.	p-value
D1 - D2	-0.22	3.23	0.41
D2 - D3	0.44	2.30	0.289
D1 - D3	0.22	2.91	0.412

S8: “The listening experience seems to be unnatural.”

Source	Mean	Std. Dev.	
D1	6.60	3.13	
D2	5.90	3.28	
D3	4.70	2.71	
Compared sources	Mean	Std. Dev.	p-value
D1 - D2	0.78	3.19	0.243
D2 - D3	0.78	1.72	0.105
D1 - D3	1.56	2.51	0.051

Table 6.8: Results of the near-field listening study.

Further test suggested

Therefore, we think that the advantage of reducing the impact of the listeners head turnings on the listening experience of the nearby source with respect to a comfortable comprehension of the content should be tested under conditions closer to the actual target application.

Confusion by unexpected change of signal

Nevertheless during the test of both setups it occurred several times while the subjects approached the sound sources, that they passed the position of the sources. In case of the sources D1, D2, and D3 at Setup 1 and source D2 at Setup 2 this lead into a sudden change of the signal of the passed source to a more muffled character, due to the incorporated monaural cue for sources in the back of the listener. Five subjects mentioned that they were confused by this unexpected change. For that reason and because this did not and can not happen in case of source D1 in Setup 2, we propose that the smooth transition between non-active and active monaural cues when passing a source, is an appropriate way to resolve this problem.

In this study the radius in which the reduction of the monaural cue was active, was dimensioned to increase the comfort of listening to source contents in the whole near-field. Therefore, we propose to evaluate whether a smaller radius would suffice, in case that further studies reveal that the comprehension of a content gets more comfortable by the reduction of monaural cues.

32 dB volume difference enables effortless comprehension

The results of Setup 1 reveal that when a subject wants to follow the content of a source, a volume difference of 26 dB between the signal of the source and the signals of the other sources is too low to enable an effortless following of the content. The study showed, that in case of increasing the difference by 6 dB to 32 dB, following the content was significantly easier. A further increase of the difference to 38 dB showed not to lead to a significantly easier following of the content of the nearby source.

Additional 6dB decrease not unnatural

Furthermore, the study revealed that an additional volume decrease of 6 dB to the general distance attenuation when stepping into the near-field of a source did not lead to a significant difference in the perceived realism of the listening experience. Therefore, we propose that this is a valuable approach to ensure an effortless comprehension of the content of nearby sources in the case, that the general distance attenuation of the more distant sources does not lead to a volume difference, that is large enough.

6.3 General Feedback and Suggestions

After the users had become familiar with our virtual audio space in the different tests, we described the actual target application of the system to them. After that, we asked them to express their opinions about using our system to experience a virtual historic coronation feast. The feedback was consistently positive. Most (9/10) subjects stated that they would like the navigation-by-ear concept and the exploration of an historic audio scenario solely by hearing in the coronation hall. Some participants (5/10) mentioned that it would be a nice way to present arid information in a compelling way. Five would enjoy the ability of walking around freely without following a designated route. Nevertheless, some (7/10) subjects suggested to hand out a map of the room containing the sound sources. Five of the subject explained, that they would like to have a map to get an overview of the existing sources, so that they would not miss one of them. Three mentioned that when being next to a source, it would be nice to have brief information about who is actually speaking about which topic and they proposed to give this hints by a map. Two subjects proposed to give this information by the screen of the guide or in a short verbal introduction when coming close to a source.

Subjects liked navigation-by-ear concept

Map wanted for different reasons

Chapter 7

Summary and Future Work

7.1 Summary and Contributions

Fast audio space prototyping environment

In the beginning we implemented a first interactive software which enables a fast and dynamic prototyping of virtual audio spaces via direct manipulation in a graphical interface. Due to the incorporation of the spatial audio synthesis API OpenAL and the connection to a tracking system, it also served as test environment of the created audio spaces.

Evaluation of OpenAL rendering

Several minutes of familiarization needed

In a first informal evaluation we tested the rendering quality of the spatial audio renderer which is available on the mobile device on which the final application should run. The conclusion of the test was, that an auralization approach which solely incorporates stereo panning to provide spatial cues, does not suffice to create an immersive audio space. Furthermore, it showed that the navigation through the virtual space solely by the auditive cues was only possible after several minutes of familiarization. In consequence we decided that we have to enhance the spatial audio rendering.

Auralization is a complex process and, if the aim is simulating the physical phenomenon close to the reality, it is com-

Implementation of three auralization approaches

putationally intensive. We implemented three different auralization approaches, Method A, Method B, and Method C, which use different workarounds and simplifications of the real physical effects to incorporate directional cues and, at the same time, do not exceed the capabilities of the target mobile device. Additionally, we incorporated a simple room acoustic based on offline filtered audio signals to create a more realistic listening experience and increase the source-distance perception of the listener. During the implementation we performed a preliminary user study and incorporated the obtained feedback by changing parameters of auralization approaches.

Auralization parameters adapted after preliminary user study

Evaluation of auralization approaches

In a final user study we evaluated our three auralization approaches in comparison to a simple stereo panning. In a conducted navigation performance test we unfortunately did not obtain reliable results. Nevertheless, the study lets us suggest that a study with more users and particularly with larger audio spaces should show more significant results.
The questionnaire which was additionally conducted, showed that Method A and Method B were preferred by the subjects with respect to the navigation in a virtual audio space. Although Method A showed a tendency to be judged the best of all methods, a small number of subjects rated it well below the average. We reasoned this to the usage of non-individual HRTFs which were incorporated in Method A. We concluded that, due to its robustness, Method B is the most appropriate auralization approach for our target application.

Method B most appropriate for our project

Two possible usability problems revealed

In one part of the preliminary user test which was performed during the implementation of the auralization methods, we presented a small coronation feast scenario to the participants. The informal interviews conducted afterwards revealed two possible usability problems with respect to the usage of a continuous virtual audio space to convey information.

Volume difference ensuring effortless listening

One mentioned problem was a reduced intelligibility of the content in case of a too low volume difference between the currently attended nearby source and the other more distant sources. We investigated this problem in a user study. A first result of the study was a volume difference value which makes sure, that it allows an effortless following of the currently attended source. A second result was, that it is possible to create this difference by increasing the volume attenuation of the more distant sources on the last few steps to the attended source, without reducing the felt realism of the listening experience.

No reduced realism by increased volume attenuation

A second problem which was uncovered by the coronation scenario test was, that strong changes in the audio signal due to head movements can be annoying in a situation when a visitor listens to the content of a source, but wants to look around freely. We performed a further user test to explore whether this could be a crucial problem for our target application and whether a reduction of the directional cues, which are incorporated into the signal of a nearby source, enhances the comfort in following a content. The study revealed that the participants did not look around while listening to a source content and in consequence the study did not prove, that the reduction of the directional cues increases the listening comfort. We supposed that the participants' behavior was influenced by the artificial test setup and that visitors of the coronation hall may behave in a different way.

Second expected problem did not occur in user study

Nevertheless the study let us suppose that the smooth transition between active and not active monaural cue in the near-field of a source is an appropriate way to reduce the confusion which was induced by the sudden changes in perceived audio signals when passing a nearby sound source.

Smooth transition to avoid confusion

Finally, we obtained some general feedback about the target application. The idea of reviving a coronation feast in the coronation hall by a continuous virtual audio space showed up to be popular.

Idea of CORONA enjoys great popularity

7.2 Future Work

Navigation performance test with larger scenarios

Of course an important part of the future work will be to repeat and eventually adapt the studies which showed no significant results due to problems in the setup or a too small number of participants. In case of the navigation performance test, a future test should incorporate larger scenarios and ensure that they are equally difficult with respect to the navigation task.

Setup closer to target application

In case of the second test concerning the listening experience in the near-fields of the sources, the test setup should be modified to be closer to the setup of the target application, i.e., it should incorporate the dialogs which will be used in the final coronation feast scenario and we propose to conduct the test in the coronation hall.

When using HRTFs in auralization it is important, that the used HRTF is close to the individual HRTF of the listener, as described in 4.2.4—"Anatomically Correct Transfer Function". It has been shown that the incorporation of an anatomically correct transfer function of the listener, outperforms an averaged HRTF with respect to the presence of the listener in a virtual environment as well as to the source localization performance [Vaeljamaee et al. [2004], Wenzel et al. [1993b]].

Test Method A with individual HRTFs

At first it should be evaluated if the advantage of an individualization of the HRTFs still holds true if it is combined with our rendering approach. In case this can be proven, we should think about how to incorporate this in our project.

Evaluate HRTF individualization approaches in context of touristic applications

Although it is unsuitable in our context to make a complete measurement of the HRTF for every visitor, there are ways to obtain an HRTF near to the exact one. In [Xu et al. [2007]] several approaches are described, e.g., several signals created with different HRTFs are presented to a listener and by a process of elimination the optimal HRTF is determined [Iwaya [2006]]. Another approach uses pictures of the listener's right and left ear to estimate the anatomical features and by that select an optimal HRTF (see figure 7.1). In [Zotkin et al. [2002b]] the authors describe such a method, which does not take longer than one minute.

A future work would be to find out which of these methods would be appropriate in the context of touristic applications, because not all visitors would accept such procedures.

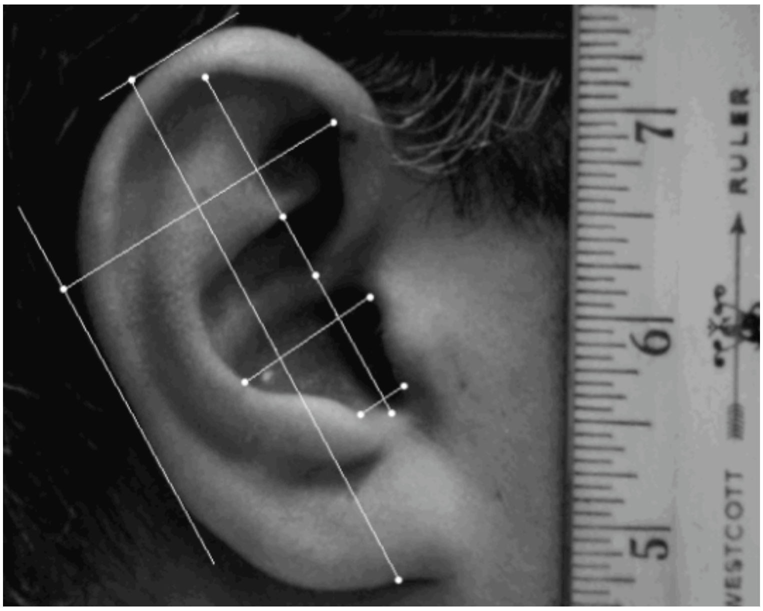


Figure 7.1: Ear with marked anatomical features for the selection of an optimal HRTF (taken from [Zotkin et al. [2002b]]).

Spatial layout of sources

The navigation performance test lets us suppose that one of the used spatial source setups increased the difficulty of the navigation task. Therefore an interesting field for future investigations would be to explore in which way the spatial layout of the sound sources influences the navigation performance and the experience in general.

Control of source playing state

An interesting question with respect to the target application, is how the playing state of the sources are controlled. Currently, it is planned for the target application to play the clips in a simple loop mode, which means that the playing state is independent from the position of the visitor. The time progression of the audio clips is synchronized for all visitors. As a result, visitors which are close to each other, e.g., companions which walk side by side, always hear the same source contents at the same time. The thought behind this is to minimize the isolation of the visitors [Aoki et al.

[2002], Stahl [2007], Heller et al. [2009]]. A question which has to be investigated is, whether it is a good tradeoff with respect to the problem that the visitors may possibly arrive at a virtual sound source, respectively at a dialog, which is already in the middle of its content.

Additional brief information about content

The user study revealed that when exploring the coronation feast scenario of the target application some visitors would desire to get a short additional information about the content of the virtual conversations and the speakers. There were also some suggestions to deliver this information, e.g., as a ground map with short explanations or the incorporation of short verbal descriptions, played when coming close to a source. A possible future work would be to investigate whether the wish of brief extra information also appears under real usage conditions, and if, to find an appropriate solution. A further point of investigation would be the influence of the additional information and its presentation style on the immersion of the visitor in the virtual audio space and whether, e.g., a map would reduce the pleasure of exploring an audio space solely by auditive cues.

Appendix A

Navigation Performance Test Scenarios

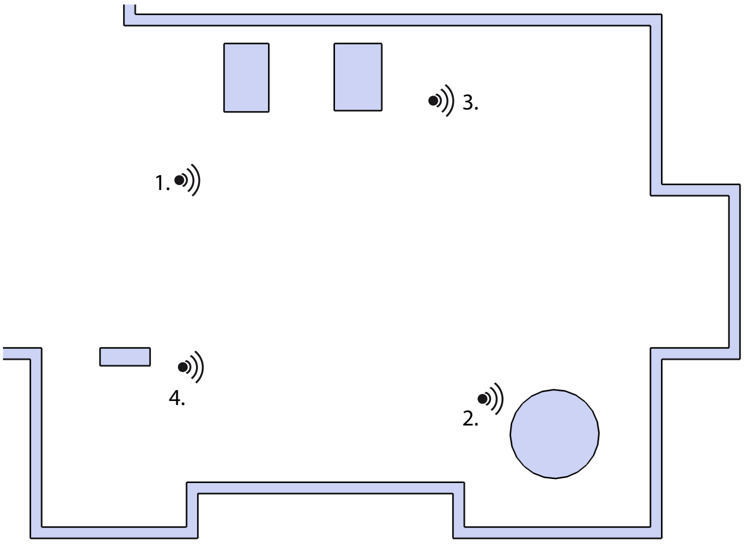


Figure A.1: Ground plan of a navigation performance test scenario with sources and the order in which they have to be approached.

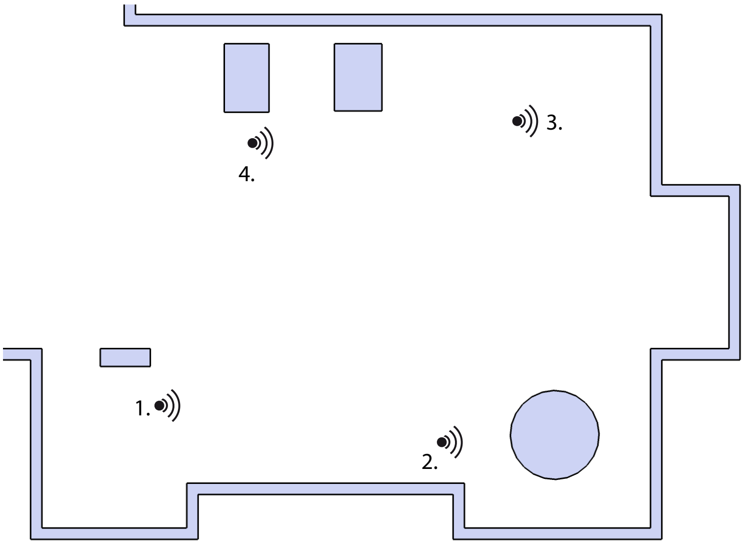


Figure A.2: Ground plan of a navigation performance test scenario with sources and the order in which they have to be approached.

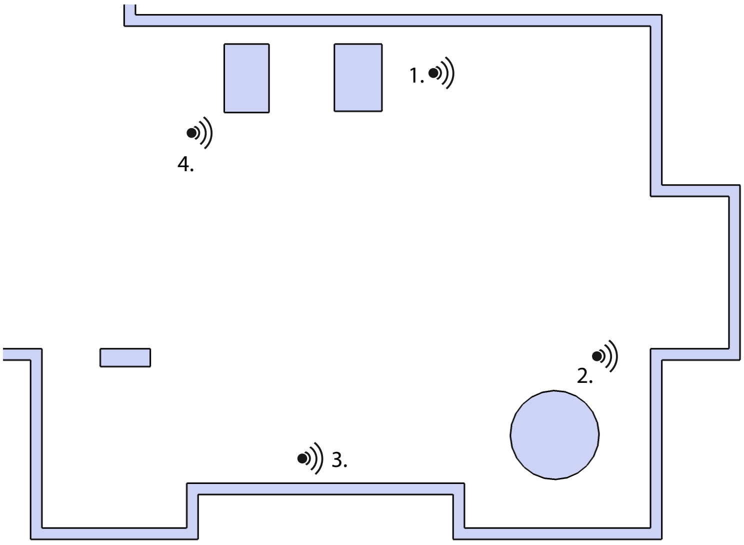


Figure A.3: Ground plan of a navigation performance test scenario with sources and the order in which they have to be approached.

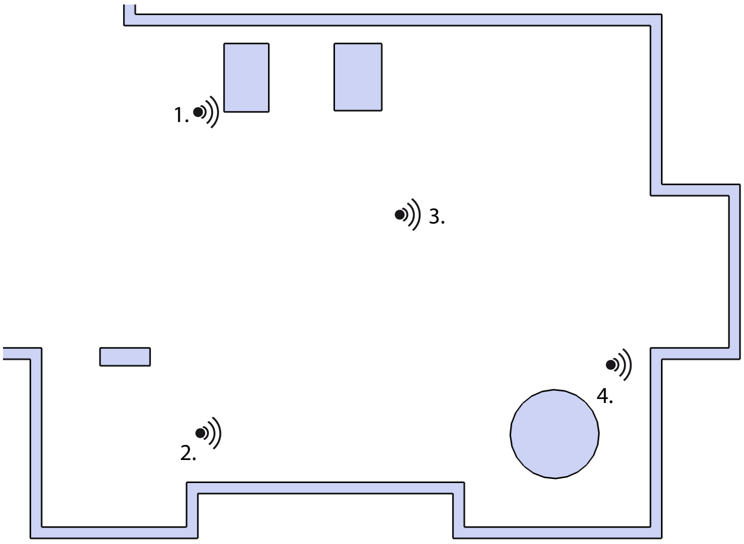


Figure A.4: Ground plan of a navigation performance test scenario with sources and the order in which they have to be approached.

Bibliography

3d audio rendering and evaluation guidelines. Technical report, 3D Working Group of the Interactive Audio Special Interest Group, 1998.

V.R. Algazi and R.O. Duda. Immersive spatial sound for mobile multimedia. In *Seventh IEEE International Symposium on Multimedia*, Dec. 2005. doi: 10.1109/ISM.2005.69.

Paul M. Aoki, Rebecca E. Grinter, Amy Hurst, Margaret H. Szymanski, James D. Thornton, and Allison Woodruff. Sotto voce: exploring the interplay of conversation and mobile audio spaces. In *CHI '02: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 431–438, New York, NY, USA, 2002. ACM. ISBN 1-58113-453-3. doi: http://doi.acm.org/10.1145/503376.503454.

Benjamin B. Bederson. Audio augmented reality: a prototype automated tour guide. In *CHI '95: Conference companion on Human factors in computing systems*, pages 210–211, New York, NY, USA, 1995. ACM. ISBN 0-89791-755-3. doi: http://doi.acm.org/10.1145/223355.223526.

D. R. Begault. Preferred sound intensity increase for sensation of half distance. *Perceptual Motor Skills*, 72:1019–1029, 1991.

J. Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization*. MiT Press, Cambridge, MA., 1983.

D.S. Brungart. Control of perceived distance in virtual audio displays. *Engineering in Medicine and Biology Society, 1998. Proceedings of the 20th Annual International Conference of the IEEE*, 3:1101–1104 vol.3, Oct-1 Nov 1998. doi: 10.1109/IEMBS.1998.747063.

Kirsten Cater, Constance Fleuriot, Richard Hull, and Josephine Reid. Experience design guidelines for creating situated mediascapes. Technical Report CSTR-06-009, Mobile and Media Systems Laboratory, October 2005. URL http://www.cs.bris.ac.uk/Publications/Papers/2000510.pdf.

E. Colin Cherry. Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*, 25(5):975–979, 1953. doi: 10.1121/1.1907229. URL http://link.aip.org/link/?JAS/25/975/1.

Steven Dow, Jaemin Lee, Christopher Oezbek, Blair Maclntyre, Jay David Bolter, and Maribeth Gandy. Exploring spatial narratives and mixed reality experiences in oakland cemetery. In *ACE '05: Proceedings of the 2005 ACM SIGCHI International Conference on Advances in computer entertainment technology*, pages 51–60, New York, NY, USA, 2005. ACM. ISBN 1-59593-110-4. doi: http://doi.acm.org/10.1145/1178477.1178484.

Richard O. Duda. *3-D Audio for HCI*. Department of Electrical Engineering Department of Electrical Engineering San Jose State University, 2000.

N. I. Durlach, A. Rigopulos, X. D. Pang, W. S. Woods, A. Kulkarni, H. S. Colburn, and E. M. Wenzel. On the externalization of auditory images. *Presence: Teleoper. Virtual Environ.*, 1(2):251–257, 1992. ISSN 1054-7460.

Gerhard Eckel. Immersive audio-augmented environments: The listen project. In *IV '01: Proceedings of the Fifth International Conference on Information Visualisation*, page 571, Washington, DC, USA, 2001. IEEE Computer Society. ISBN 0-7695-1195-3.

Richard Etter and Marcus Specht. Melodious walkabout - implicit navigation with contextualized personal audio contents. In *In Adj. Proc. Pervasive Computing*, page 43. Technology, 2005.

Mark B. Gardner. Distance estimation of 0[degree] or apparent 0[degree]-oriented speech signals in anechoic space. *The Journal of the Acoustical Society of America*, 45(1):47–53, 1969. doi: 10.1121/1.1911372. URL http://link.aip.org/link/?JAS/45/47/1.

Joachim Gossmann and Marcus Specht. Location models for augmented environments. *Personal Ubiquitous Comput.*, 6(5-6):334–340, 2002. ISSN 1617-4909. doi: http://dx.doi.org/10.1007/s007790200038.

William M. Hartmann. How we localize sound. *Physics Today*, Nov 1999. URL `http://www.aip.org/pt/nov99/locsound.html`.

M. L. Hawley, R. Y. Litovsky, and J. F. Culling. The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer. *Acoustical Society of America Journal*, 115:833–843, February 2004. doi: 10.1121/1.1639908.

Florian Heller. Corona - realizing an interactive experience in visually untouchable rooms using continuous virtual audio spaces. Master's thesis, RWTH Aachen University, December 2008.

Florian Heller, Thomas Knott, Malte Weiss, and Jan Borchers. Multi-user interaction in virtual audio spaces. In *CHI EA '09: Proceedings of the 27th international conference extended abstracts on Human factors in computing systems*, pages 4489–4494, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-247-4. doi: http://doi.acm.org/10.1145/1520340.1520688.

Simon Holland, David R. Morse, and Henrik Gedenryd. Audiogps: Spatial audio navigation with a minimal attention interface. *Personal Ubiquitous Comput.*, 6(4):253–259, 2002. ISSN 1617-4909. doi: http://dx.doi.org/10.1007/s007790200025.

M.G. Hood. Staying away: Why people choose not to visit museums. *Museum News*, 61(4):50–57, 1983.

Yukio Iwaya. Individualization of head-related transfer functions with tournament-style listening test: Listening with others ears. *Acoustical Science and Technology*, 27(6):340–343, 2006.

J. M. Loomis, C. Hebert, and J. G. Cicinelli. Active localization of virtual sounds. In *NASA. Ames Research Center, Human Machine Interfaces for Teleoperators and Virtual Environments p 134 (SEE N95-14013 03-54)*, pages 134–+, June 1991.

Accoustiguide Maryanne Leigh. The next generation of guiding technology maryanne leigh, accoustiguide. In *BUILDING BLOCKS Conference 2007*, volume 15. Interpretation Australia Association, November 2007.

Luca Mion, Federico Avanzini, Bruno Mantel, Benoit Bardy, and Thomas A. Stoffregen. Real-time auditory-visual distance rendering for a virtual reaching task. In *VRST '07: Proceedings of the 2007 ACM symposium on Virtual reality software and technology*, pages 179–182, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-863-3. doi: http://doi.acm.org/10.1145/1315184.1315217.

Nancy Proctor and Chris Tellis. The state of the art in museum handhelds in 2003. In *Museums and the Web 2003*, Charlotte, North Carolina, USA, March 2003. Archives & Museum Informatics.

Josephine Reid, Richard Hull, Kirsten Cater, and Benjamin Clayton. Riot! 1831: The design of a location based audio drama. In *Proceedings of UK-UbiNet 2004*, pages 1733–1736. UK-UbiNet, October 2004. URL `http://www.cs.bris.ac.uk/Publications/Papers/2000261.pdf`.

Josephine Reid, Erik Geelhoed, Richard Hull, Kirsten Cater, and Ben Clayton. Parallel worlds: immersion in location-based experiences. In *CHI '05: extended abstracts on Human factors in computing systems*, pages 1733–1736, New York, NY, USA, 2005a. ACM. ISBN 1-59593-002-7. doi: http://doi.acm.org/10.1145/1056808.1057009.

Josephine Reid, Richard Hull, Kirsten Cater, and Constance Fleuriot. Magic moments in situated mediascapes. In *ACE '05: Proceedings of the 2005 ACM SIGCHI International Conference on Advances in computer entertainment technology*, pages 290–293, New York, NY, USA, 2005b. ACM. ISBN 1-59593-110-4. doi: http://doi.acm.org/10.1145/1178477.1178529.

Wakkary Ron, Newby Kenneth, Hatala Marek, Evernden Dale, Droumeva Milena, and Fraser Simon. Interactive audio content: An approach to audio content for a dynamic museum experience through augmented audio reality and adaptive information retrieval. In *Museums* *and the Web 2004*, Toronto, Ontario, Canada, April 2004. Archives & Museums Informatics.

B. G. Shinn-Cunningham. Distance cues for virtual auditory space. Invited Paper: IEEE-PCM Special Session on Virtual Auditory Space, 2000.

J. M. Speigle and J. M. Loomis. Auditory distance perception by translating observers. *Virtual Reality, 1993. Proceedings., IEEE 1993 Symposium on Research Frontiers in*, pages 92–99, 1993.

Walter Spieth, James F. Curtis, and John C. Webster. Responding to one of two simultaneous messages. *The Journal of the Acoustical Society of America*, 26(3):391–396, 1954. doi: 10.1121/1.1907347. URL `http://link.aip.org/link/?JAS/26/391/1`.

Christoph Stahl. The roaring navigator: a group guide for the zoo with shared auditory landmark display. In *MobileHCI '07: Proceedings of the 9th international conference on Human computer interaction with mobile devices and services*, pages 383–386, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-862-6. doi: http://doi.acm.org/10.1145/1377999.1378042.

Kay M. Stanney, editor. *Handbook of Virtual Environments: Design, Implementation, and Applications.* Lawrence Erlbaum Associates, Mahwah, NJ, 2002.

J.W. Strutt. On our perception of sound direction. *Phil. Mag.*, 13:214–232, 1907.

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

Aleksander Vaeljamaee, Pontus Larsson, and Daniel Vaestfjaell. Auditory presence, individualized head-related transfer functions, and illusory ego-motion in virtual environments. In *7th Annual Workshop Presence*, 2004.

Michael Vorländer. *Auralization*. Springer-Verlag, Berlin, 2007.

Ron Wakkary and Marek Hatala. Situated play in a tangible interface and adaptive audio museum guide. *Personal Ubiquitous Comput.*, 11(3):171–191, 2007. ISSN 1617-4909. doi: http://dx.doi.org/10.1007/s00779-006-0101-8.

Nigel Warren, Matt Jones, Steve Jones, and David Bainbridge. Navigation via continuously adapted music. In *CHI '05: CHI '05 extended abstracts on Human factors in computing systems*, pages 1849–1852, New York, NY, USA, 2005. ACM. ISBN 1-59593-002-7. doi: http://doi.acm.org/10.1145/1056808.1057038.

E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman. Localization using nonindividualized head-related transfer functions. *Acoustical Society of America Journal*, 94:111–123, July 1993a. doi: 10.1121/1.407089.

E. M. Wenzel, M. Arruda, D.J. Kistler, and F.L. Wightman. Localization using nonindividualized head-related transfer functions. *Acoustical Society of America Journal*, 94:111–123, July 1993b. doi: 10.1121/1.407089.

Song Xu, Zhizhong Li, and Gaviriel Salvendy. *Individualization of Head-Related Transfer Function for Three-Dimensional Virtual Auditory Display: A Review*, pages 397–407. Lecture Notes in Computer Science. Springer Berlin, 2007.

Dmitry Zotkin, Ramani Duraiswami, and Larry S. Davis. Rendering localized spatial audio in a virtual auditory space. 2002a. URL `http://hdl.handle.net/1903/1190`.

Dmitry N. Zotkin, Ramani Duraiswami, Larry S. Davis, Ankur Mohan, and Vikas Raykar. Virtual audio system customization using visual matching of ear parameters. In *ICPR '02: Proceedings of the 16 th International Conference on Pattern Recognition (ICPR'02) Volume 3*, page 31003, Washington, DC, USA, 2002b. IEEE Computer Society. ISBN 0-7695-1695-X.

Index

Typeset June 15, 2009