Department of Media Technology

Interaction with eyes-free and gestural interfaces

Raine Kajastila

Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 22/2013

Interaction with eyes-free and gestural interfaces

Raine Kajastila

A doctoral dissertation completed for the degree of Doctor of Science in Technology to be defended, with the permission of the Aalto University School of Science, at a public examination held at the lecture hall AS1 of the Aalto University School of Science (Espoo, Finland) on 8th of February 2013 at 12.

Aalto University
School of Science
Department of Media Technology

Supervising professor
Associate Professor Tapio Lokki

Thesis advisor
Associate Professor Tapio Lokki

Preliminary examiners
Associate Professor Bruce N. Walker,
Georgia Institute of Technology in Atlanta,
the United States of America

Assistant Professor Federico Avanzini,
University of Padova,
Italy

Opponents
Professor Stephen Brewster,
University of Glasgow,
United Kingdom

Aalto University publication series
DOCTORAL DISSERTATIONS 22/2013

ISBN 978-952-60-5002-7 (printed)
ISBN 978-952-60-5003-4 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-5003-4

Unigrafia Oy
Helsinki 2013

Finland

Publication orders (printed book):
http://lib.tkk.fi/Diss/

441 697
Printed matter

Aalto University

Abstract

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Author
Raine Kajastila

Name of the doctoral dissertation
Interaction with eyes-free and gestural interfaces

Publisher School of Science

Unit Department of Media Technology

Series Aalto University publication series DOCTORAL DISSERTATIONS 22/2013

Field of research Media technology

Manuscript submitted 11 September 2012 **Date of the defence** 8 February 2013

Permission to publish granted (date) 17 December 2012 **Language** English

☐ **Monograph** ☒ **Article dissertation (summary + original articles)**

Abstract

Eyes-free interaction aims to control devices without the need to look at them. This is especially useful while driving, walking on a bustling street, or in other situations when looking at a display would be dangerous, inconvenient or restricted. Hand gestures and feedback with sound offer an eyes-free alternative to visual displays, and this thesis studies using them with devices and the surrounding environment.

In this thesis work, advanced circular auditory menus and three parallel control methods for using them were developed. Essentially, the thesis work concentrated on a circular interaction metaphor in auditory menus, in which the gesture was mapped directly to the position in the menu. The introduced control methods and auditory menu properties were tested with user experiments, and a mobile application integrating auditory and visual menus was built.

The three gestural control methods to control circular auditory menus included accelerometer-based, touch screen-based, and camera-based interaction. All control methods were proven accurate and fast enough for efficient eyes-free use. Additionally, the same control methods were used in both visual and auditory domains, which facilitates switching to eyes-free use when needed and may also improve the accessibility of the interface for visually impaired users. Results of user experiments showed that the introduced visual and auditory menu design was easy and intuitive to learn without extensive training. Furthermore, a solution for eyes-free access to large menus was proposed, and user experiments indicated that dynamic menu item placement is efficient, accurate, and allowed the use of large menus.

This thesis also investigated the use of auditory displays and gesture interfaces in performing arts. The perceived shape and size of a space can be changed by applying different reverberation times in different directions using multiple reverberation systems. Implementing a reverberation system and a test setup for subjective evaluation validated this. The implemented reverberation system has been utilized in live opera performances and to enhance lecture room acoustics. The use of gesture control is explored in an experimental opera production in which the performers controlled an audiovisual virtual stage live. The live interaction was useful when user controlled media was directly mapped onto gestures and when detailed nuances of movement were hard for a technician controlling the media to follow.

Keywords Eyes-free, gesture interaction, auditory menus, auditory display, spatial sound, interface, HCI, performing arts

ISBN (printed) 978-952-60-5002-7 **ISBN (pdf)** 978-952-60-5003-4

ISSN-L 1799-4934 **ISSN (printed)** 1799-4934 **ISSN (pdf)** 1799-4942

Location of publisher Espoo **Location of printing** Helsinki **Year** 2013

Pages 170 **urn** http://urn.fi/URN:ISBN:978-952-60-5003-4

A? Aalto-yliopisto

Tiivistelmä

Aalto-yliopisto, PL 11000, 00076 Aalto www.aalto.fi

Tekijä
Raine Kajastila

Väitöskirjan nimi
Vuorovaikutus ele- ja äänikäyttöliittymillä

Julkaisija Perustieteiden korkeakoulu

Yksikkö Mediatekniikan laitos

Sarja Aalto University publication series DOCTORAL DISSERTATIONS 22/2013

Tutkimusala Mediatekniikka

Käsikirjoituksen pvm 11.09.2012 **Väitöspäivä** 08.02.2013

Julkaisuluvan myöntämispäivä 17.12.2012 **Kieli** Englanti

☐ **Monografia** ☒ **Yhdistelmäväitöskirja (yhteenveto-osa + erillisartikkelit)**

Tiivistelmä

Käyttöliittymiä joita ei tarvitse katsoa kutsutaan eyes-free käyttöliittymiksi. Ne mahdollistavat turvallisemman laitteiden käytön tilanteissa, joissa näytön katsominen on hankalaa, vaarallista tai mahdotonta, kuten esimerkiksi autoa ajaessa tai kävellessä vilkkaalla kadulla. Käden eleillä ja äänipalautteella toteutetut eyes-free käyttöliittymät ovat vaihtoehto visuaalisille näytöille ja tässä väitöskirjassa esitellään niiden käyttämistä laitteiden ohjaukseen ja vuorovaikutukseen ympäristön kanssa.

Tässä väitöstyössä tutkittiin edistyksellisiä äänivalikkoja, joita ohjattiin kolmella rinnakkaisella ohjausmenetelmällä. Käytetyt ohjausmenetelmät perustuivat käden eleisiin, joita seurattiin kiihtyvyysantureiden, kosketusnäytön tai kameran avulla. Pääasiallisesti väitöstyössä keskityttiin ympyränmuotoisen vuorovaikutusmetaforan käyttöön äänivalikoissa, joissa käden liike määritti paikan ympyrävalikossa. Esitettyjen ohjausmenetelmien ja valikkoratkaisuiden toimivuus arvioitiin käyttäjäkokeilla ja lisäksi toteutettiin älypuhelinsovellus, joka yhdisti ääni- sekä visuaalisten valikkojen toiminnallisuuden.

Kolme kehitettyä äänivalikkojen ohjausmenetelmää osoitettiin käyttäjäkokeilla tarkoiksi ja riittävän nopeiksi käyttöön äänipalautteen kanssa. Lisäksi samat ohjausmenetelmät soveltuivat äänivalikon kanssa yhtenevän visuaalisen valikon ohjaamiseen, mikä voi helpottaa äänivalikon käyttöönottoa sekä parantaa käyttöliittymien yleistä soveltuvuutta näkövammaisille. Käyttäjätestin tulokset osoittivat, että audiovisuaalinen valikkoratkaisu oli helppo sekä intuitiivinen oppia ilman pitkää harjoittelua. Äänivalikkoihin kehitettiin myös uusi tapa selata isoja valikkorakenteita, joka todistettiin käyttäjäkokeilla tehokkaaksi.

Tässä väitöskirjassa tutkittiin myös äänipalautetta ja elekäyttöliittymiä esittävässä taiteessa. Ympäröivän tilan havaittua kokoa ja muotoa säädeltiin muuttamalla sähköisesti seinistä tulevia äänten heijastuksia kaikulaitteilla. Tämä osoitettiin mahdolliseksi käyttäjäkokeilla ja varta vasten rakennetulla kaikulaitejärjestelmällä, jota käytettiin myös oopperatuotannossa sekä luentosalin akustiikan parantamiseen. Oopperatuotannon osana tutkittiin myös eleohjauksen ja audiovisuaalisen lavastuksen vuorovaikutusta. Esiintyjien käyttämä eleohjaus oli hyödyllistä etenkin improvisoitavissa kohtauksissa tai kun ulkopuolisen ohjaajan oli hankala seurata näyttelijöiden liikkeiden vivahteita.

Avainsanat Elekäyttöliittymä, äänivalikot, tilaääni, käyttöliittymä, eyes-free, taide

ISBN (painettu) 978-952-60-5002-7 **ISBN (pdf)** 978-952-60-5003-4

ISSN-L 1799-4934 **ISSN (painettu)** 1799-4934 **ISSN (pdf)** 1799-4942

Julkaisupaikka Espoo **Painopaikka** Helsinki **Vuosi** 2013

Sivumäärä 170 **urn** http://urn.fi/URN:ISBN:978-952-60-5003-4

Preface

The research work for this thesis has been carried out in the Telecommunications Software and Multimedia Laboratory (TML) and in the Department of Media Technology in Helsinki University of Technology during 2006–2009, and in Aalto University during 2010–2012. Furthermore, part of the research has been made in collaboration with Helsinki Institute for Information Technology (HIIT), Theatre Academy Helsinki (Teak), and Sibelius Academy (Siba). The research leading to these results has also received funding from the Academy of Finland, project no. [119092], the European Research Council under the European Community's Seventh Framefork Programme / ERC grant agreement no. [203636], and the CALLAS project (IST-034800) under the European Community's Sixth Framefork Programme.

I want to express my full gratitude to my instructor and supervisor Prof. Tapio Lokki for letting me find my own research topics and guiding the research, for inspiring innovation sessions and support throughout all the funky entrepreneurial business in between and then finally welcoming me back for the finalization of this thesis. I wish to thank Prof. Lauri Savioja for hiring me in the first place and for guidance during these years, and also Prof. Tapio Takala for the possibility to work with the inspiring opera production.

The department of Media Technology (and formerly TML) has always had an inspiring atmosphere. It has been a pleasure to work with you all! Special thanks goes to all co-workers at PhdVirta, to all who participated in the inspiring corner-room coffee discussions, and to the wonderful staff. I'm also grateful to all who were involved with the Koala project and who contributed to the building of the most fungestorous mobile app.

I wish to thank everybody who participated in the virtual opera production and contributed to the fruitful and interdisciplinary collaboration between Helsinki University of Technology, Helsinki Institute for Information Technology, Theater Academy, and Sibelius Academy. I'm also thankful to the people at Finnish Federation of the Visually Impaired for testing the user interface prototypes and for the valuable discussions with them.

I would like to thank the pre-examiners of this thesis, Dr. Bruce N. Walker and Dr. Federico Avanzini, for valuable feedback and constructive comments for the manuscript. A special thanks also goes to Luis Costa for proof-reading this manuscript.

I'm grateful to my family and friends for all the support and friendship they have given me. Finally, my most sincere thanks go to Jaana for endless support, love and encouragement and to Eevo for reminding me everyday what are the most important things in life.

Espoo, January 14, 2013,

Raine Kajastila

Contents

List of Publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

I Tapio Lokki, Raine Kajastila and Tapio Takala. Virtual acoustic spaces with multiple reverberation enhancement systems. In *Proceedings of the AES 30th international conference*, CD-ROM Proceedings, Saariselkä, Finland, March 15–17 2007.

II Raine Kajastila and Tapio Takala. Interaction in digitally augmented opera. In *Proceedings of international conference on digital arts*, pp. 216–219, Porto, Portugal, November 7–8 2008.

III Raine Kajastila and Tapio Lokki. A gesture-based and eyes-free control method for mobile devices. In *Proceedings of the 27th international conference extended abstracts on human factors in computing systems (CHI '09)*, pp. 3559–3564, Boston, MA, USA, April 4–9 2009.

IV Raine Kajastila and Tapio Lokki. Eyes-free methods for accessing large auditory menus. In *Proceedings of 16th international conference on auditory display (ICAD 2010)*, pp. 223–230, Washington, D.C, USA, June 9–15 2010.

V Raine Kajastila and Tapio Lokki. Eyes-free interaction with free-hand gestures and auditory menus. *International journal of human-computer studies*, Accepted for publication (In press), 14 pages, November 2012.

VI Raine Kajastila and Tapio Lokki. Combining auditory and visual menus. *Acta acustica united with acustica*, vol. 98, no. 6, pp. 945-956, 2012.

Author's Contribution

Publication I: "Virtual acoustic spaces with multiple reverberation enhancement systems"

This publication presents a way to modify of the size and shape of an auditory perceived space with multiple reverberation enhancement systems. The acoustic response is generated with a time-variant reverberation algorithm which prevents acoustic feedback. A subjective test indicated that by applying different reverberation times in different directions, the perceived shape and size of the space can be changed. The prototyped system is used in live performances and in education.

The present author implemented the reverberation system. The experiment was set up and conducted together with Prof. Tapio Lokki. Prof. Tapio Lokki wrote most of the article, and the present author and Prof. Tapio Takala aided in the writing process.

Publication II: "Interaction in digitally augmented opera"

An experimental opera production is described, where digitally augmented content was used interactively during the performance. Projected graphics and spatial sounds were designed to support the story. The animated 3-D graphics acted either as a virtual stage, a narrative element, or a reflection of the thoughts of a character. The special effects were partly in the performers' direct control via accelerometers, thus allowing eyes-free use, enabling natural timing and giving more freedom for artistic expression.

The present author designed and implemented the visual and acoustic virtual stage environment used in the opera and wrote 90% of the article.

Publication III: "A gesture-based and eyes-free control method for mobile devices"

A novel interaction method for eyes-free control of a mobile device is introduced. A spherical auditory menu and feedback are provided using speech, and rendered with spatial sound. A gestural pointing interface, multiple menu configurations, and their implementation details are presented. Evaluation results suggest that fast and accurate selection of menu items is possible without visual feedback.

The present author invented the gestural control method, implemented and conducted all of the experiments and wrote 90% of the article. Half of the statistical analysis of the results was done by Prof. Tapio Lokki.

Publication IV: "Eyes-free methods for accessing large auditory menus"

The effectiveness of the introduced gestural and touch screen interaction is compared to a traditional visual interface when accessing large menus. A new browsing method with dynamically adjustable menu item positions is used with large menus, and evaluation results show that moderately fast and accurate browsing of large menus is possible without visual feedback.

The present author invented the eyes-free method for browsing long lists with gestural control, implemented and conducted all of the experiments, and wrote 90% of the article. Half of the statistical analysis of the results was done by Prof. Tapio Lokki.

Publication V: " Eyes-free interaction with free-hand gestures and auditory menus"

A novel free-hand gesture interaction with camera-based tracking and auditory menus is presented. This publication describes a user study where tested control methods included the free-hand gesture interaction with camera-based tracking and touch screen interaction with a tablet. The results show that even with the participant's full attention on the task, the performance and accuracy of the auditory interface is comparable to or even slightly better than the visual interface, when the two interfaces

are controlled with free-hand gestures.

The present author invented the free-hand control method, implemented and conducted all of the experiments, and wrote 90% of the article. Half of the statistical analysis of the results was done by Prof. Tapio Lokki.

Publication VI: "Combining auditory and visual menus"

This publication presents an interoperable control interface for the auditory and visual domain. The same control logic for both visual and auditory domains can facilitate switching to eyes-free use when needed. The novel interface paradigm is explained with the Funkyplayer application that allows eyes-free touch screen and gesture access to a music collection on a mobile phone. The results of the user experiment show that auditory and visual menus with the same control logic can provide a fast, usable, and intuitive interface to control devices.

The present author designed 90% of the Funkyplayer application, conducted all of the experiments, and wrote 90% of the article. Half of the statistical analysis of the results was done by Prof. Tapio Lokki. The programming and the final visual appearance was the joint effort of the KOALA research group.

List of Abbreviations

3D	three-dimensional
AD	analog-to-digital
DA	digital-to-analog
EMG	electromyography
GUI	graphical user interface
FIR	finite Impulse Response
HRTF	head related transfer function
ILD	interaural level difference
iOS	mobile operating system for iPhone
ITD	interaural timel difference
IVR	interactive voice response system
IVS	in vehicle systems
OSX	operating system for Macintosh computers
PC	personal computer
PD	pure data
RES	reverberation enhancement system
RT	reverberation time
VBAP	vector base amplitude panning
WIMP	windows, icons, menus, pointer

1. Introduction

This thesis studies new methods for eyes-free interaction, especially concentrating on gestural interaction with auditory menus and interacting with performance spaces.

Auditory interfaces can bring better usability in situations where eyes-free operation is necessary [15]. Such cases include the competition of visual attention, absence or limitations of visual display, or reduction of battery life [138]. With proper design, an auditory interface can be even more effective than its visual counterparts [138]. Auditory interfaces can overcome visual interfaces especially when a second task, such as driving, competes for the attention of a user. Furthermore, they are important as assistive technology for visually impaired users. Screen reading applications make the reading of text possible and auditory menus are used to replace the visual menus in computer programs.

Auditory interfaces are becoming more common in everyday life. For example, Apple has introduced the iPod shuffle [47], which gives feedback to user using synthesized speech. A device, such as the iPod shuffle, without a visual display, can be used eyes-free, is inexpensive to manufacture, and has low energy consumption.

A typical way of controlling a device is to first reach for a specific controller and then operate the device while looking at a display. This can draw the user's focus to the device for a long time. Often, the interaction with the device should distract the user from the task at hand as little as possible, simultaneously enabling efficient control. Gesture interaction can offer a natural control interface and, using auditory feedback, can free the eyes for another task. However, using audio as feedback is often overlooked when designing novel interactions for gesture control.

The performances and spaces can also benefit from eyes-free and gesture interaction. Actors concentrate deeply on their performance, and enabling

natural and eyes-free control can help them. The performer-based interaction is useful when user controlled media is directly mapped to gestures and when detailed nuances of movement are hard for a technician controlling the media to follow. Acoustics of the room can be modified and the interaction with the room's electronically enhanced sound environment may be used to change the perceived auditory shape of the performance room, support the narrative of a performance, or even create an entity with instrumental music. Reverberation systems can also be used to electronically augment a rehearsal room to better match the acoustics of the performance hall [69], thus giving the musicians the feeling of playing on the actual stage.

1.1 Scope of this thesis

This thesis studies approaches for eyes-free and auditory interaction with devices and spaces. Eyes-free interaction is a relatively wide topic, and this thesis focuses on interaction where feedback to the user is mainly given with audio. Accordingly, the control methods are constrained mostly to hand gestures.

The scope of this thesis and related topics are explained in Figure 1.1. The contribution and the relation between the publications is the following:

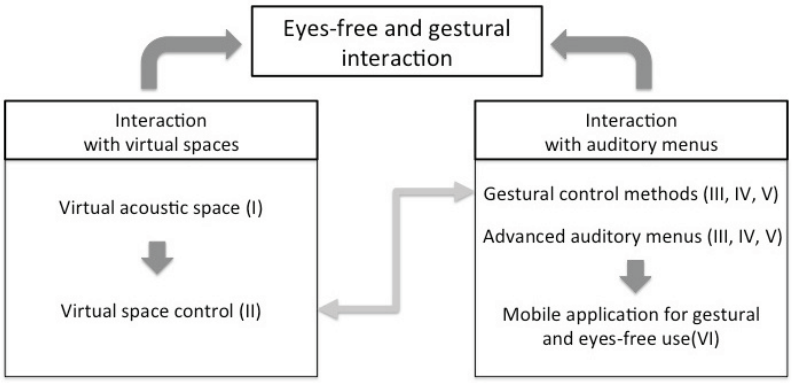

Figure 1.1. Scope of the thesis and the concepts it covers. The publications discussing the topics are indicated with Roman numerals in this figure.

Interaction with virtual spaces

Virtual acoustic space

A method to modify the size and shape of a perceived auditory space with multiple reverberation enhancement systems is presented. Publication I describes an experiment where subjects interacted with the space and the reverberation system. The results indicate that applying different reverberation time in different directions can change the perceived shape and size of the space. The built reverberation system was also used to augment a live performance described in Publication II.

Virtual space control

Augmenting a live performance with a gesture controlled virtual stage and spatial sound is explored. In the experimental opera production described in Publication II, the animated 3-D graphics and sounds act as a virtual stage, a narrative element, or a reflection of the thoughts of a character. The special effects are partly under the performers' direct control, which allows eyes-free use, natural timing, and gives more freedom for artistic expression. Accelerometers were used to captue the performers' gestures, and a need for gesture controlled auditory interaction inspired and led to the work described in Publication III.

Interaction with auditory menus

Gestural control methods

Three different gestural control methods for auditory menus are explained and their performance has been tested in user studies. Publications III, IV, and V describe studies of accelerometer-based, touch screen-based and camera-based control methods, which all apply the same circular control metaphor for controlling auditory menus.

Advanced auditory menus

Techniques for advanced auditory menus are proposed. Publications III, IV, and VI include improvements to auditory menus that enable faster or easier browsing and selection of auditory menu items.

Mobile application for gestural and eyes-free use

A detailed description of created mobile application using interoperable auditory and visual menus is presented. The application presented in Publication VI is an example of how eyes-free interaction can be used in realistic settings. Furthermore, the application shows how the auditory menu and its visual counterpart can share the same efficient control logic, and it also incorporates many features from previous publications.

1.2 Organization of this thesis

This thesis consists of six publications and related background information that is organized in the following order. Chapter 2 briefly discusses the basic concepts of eyes-free interaction, hand gestures, and spatial audio. The related research on the topics of this thesis is listed in Chapter 3. Chapter 4 reviews Publications I and II and discusses interaction with a surrounding space. Chapter 5 presents the research done in Publications III, IV, V, and VI, which introduce advanced circular auditory menus and three parallel control methods for using them. Finally, in Chapter 6, the contributions of the thesis are summarized.

2. Background

This chapter gives background information on the topics presented in this thesis. First, a general reasoning behind the need for eyes-free interaction research is given. Second, an overview on the categorization of hand gesture interaction and the kind of gestures studied in this thesis follows. Finally, the chapter ends with a short introduction to how spatial sound is localized by humans and how spatial sound sources benefit applications, such as auditory menus.

2.1 Eyes-free interaction

Work on auditory [15, 12, 118] and haptic [86, 89] displays have used the term eyes-free, referring to the fact that the state of some system can be controlled or monitored without visual attention. Eyes-free interaction can take many forms, and in some situations it can be even faster than the visual counterparts [138]. The reasons why eyes-free interaction is needed or desirable [80, 138, 136] are discussed below:

Competition for visual attention: The obvious reason for eyes-free use is the need for interfaces which do not compete with visual attention. Eyes-free interfaces can efficiently keep visual concentration on the road while diving [107] or when walking in the streets of a bustling city [118]. In these situations, focusing visual attention on a device (even for a short time) can be a risk.

User disability: Eyes-free interfaces are important as assistive technology for visually impaired users. Screen readers, auditory menus [138], and eyes-free text entry [39] enable access to computers and digital information. Implementing completely different interfaces

for sighted and visually impaired, which engage different sensory modalities, is also justified [133].

Concentration: In many situations, the visual display can disrupt the user's concentration on the task at hand. Eyes-free and continuous monitoring systems in operating rooms can help medical staff to maintain high levels of awareness of patients' state while concentrating on other tasks [130]. Also in sports, eyes-free interfaces can be used to give real-time feedback when the athlete is fully concentrating on the activity itself [110]. Likewise in the performing arts, concentration in essential. For example, adjusting visual knobs and sliders may require too much concentration and effort from a performer. By using gestures and the human body, the performer may control various effects at once with less effort [135].

Environmental restrictions: Bright light can make visual displays unreadable, and equipment used in such conditions can benefit from eyes-free interaction. Also, there are situations where the light produced by the visual display is not desired, such as while developing a photography film in a dark room or avoiding being seen by others in special police or military operations in the dark. The surrounding environment also affects the requirements for the feedback modality, and audio can be useless in noisy environments as can tactile feedback during a bouncy car ride [45].

Absence or limitations of visual display: Ubiquitous computing hides technology, and devices get smaller and smaller. As the size of devices decreases, some devices intentionally lack a display [47]. Size of a device can also cause the visual display to be too small for efficient use or, on the other hand, size of a device cannot be reduced without visual display becoming unusable [61].

Reduction of cost and battery life: Efficient eyes-free interaction could enable smaller and cheaper mobile devices. Small devices (e.g., iPod shuffle [47]) can be cheap to manufacture and also their power consumption can be kept low without having large displays.

Inconvenience: Sometimes digging a phone out of the pocket to check for new received messages is just inconvenient and thus simple auditory notifications are widely used. More sophisticated vibrotactile and auditory cues introduced in Shoogle [132] can provide even more

information. The Shoogle prototype enables eyes-free and a natural way to check the number of messages in the inbox by using a metaphor of bouncing balls inside a shaken device or by replacing the visual cue of the battery life indicator with the sound of liquid in a bottle.

Social acceptability: Even though the use of mobile devices is common in everyday life, in some social context operating a device while interacting socially with others is impolite. Eyes-free interfaces can offer unnoticeable operation of a device (e.g., checking an urgent message) during lectures, meetings, or other social interaction. Everyday objects can be used to perform socially acceptable and unnoticeable interactions such as using a ring as an input device [2].

Privacy: A visual screen might be a security risk, because its contents are visible to somebody peeking over the shoulder. Sometimes there is even need for undetectable communication even when surrounded by people. This kind of intimate interfaces can be controlled, for example, with subtle muscle gestures detected with an armband leveraging Electromyography (EMG) enabling device control which is difficult for other to detect [24] or even using devices implanted under the skin [46].

Mobility: Mobile situations often sum up the above reasons. Changes in environmental and social situations can require visual attention to be needed elsewhere. An auditory display can also be a good option for persons to whom reading in motion may induce motion sickness. Mobile situations have also special needs such as navigation. Eyes-free navigation aids range from common navigators providing turn-by-turn instructions with speech to providing a sense of direction with a vibrotactile compass [89].

Natural interaction with sound and spaces: Sound is used in natural interaction with devices. The subtle and pleasant sounds of a coffee machine indicate when coffee is ready or the continuous sound feedback from a car's motor prompts when to switch gears. Sonification (the use of non-speech audio to convey information [64]) can be used to naturally augment the interaction with the environment and as an eyes-free alternative or complement to visual information. Sounds can be also used to augment the surrounding space, e.g., to

better match the acoustics of a rehearse hall to a performance hall [69].

This thesis focuses on interaction with auditory displays that can be used to gain the above-mentioned benefits of the eyes-free interaction. The input methods are constrained mostly to hand gestures that are discussed further in the next section. Tactile feedback is commonly used to accompany visual and auditory menus and it can improve pointing interactions [1] and make touchscreen typing faster and more accurate [14]. Tactile feedback and using it with the auditory menus is research topic on it's own and thus it is out of the scope of this thesis.

Speech recognition as an eyes-free input method is viable solution [77]. It is also gaining publicity, in particular after the introduction of commercial products like Voice Actions for Android [121] and Siri on iPhone [106]. With speech recognition, the voice can be used to command devices to do specific actions. Eyes-free speech recognition interfaces are mainly command-oriented and, for example, eyes-free browsing for a long list of artists and selecting a song that fits one's mood may be harder with speech recognition. Speech recognition is still inaccurate mainly because of language and dialect barriers and can also be unusable in noisy environments. Furthermore, people might want to keep their privacy and prefer not to talk to their phone in public. However, speech recognition is outside the scope of this thesis.

2.2 Hand gesture interaction

Gesture interaction can be categorized in many ways, and there is no single definition for the term hand gesture. Hand gestures are used as a part of communication and one example is classifying them based on how they are accompanied with speech [98]:

Symbolic gestures

Have a single meaning such as the peace sign or individual gestures in sign language.

Deictic gestures

Directs the listener's attention by pointing or other means.

Iconic gestures

Describe the physical properties of an object or how it moves.

Pantomimic gestures

Describe how an object is used and held in hand.

They also can be classified according to their function [19, 18]:

Semiotic (communication)

Communication of information towards the environment.

Ergotic (manipulation, creation)

Material action, modification and transformation of the environment.

Epistemic (touching, feeling)

Perception of the environment.

Also in human computer interaction the gestures have been categorized generally [21, 57] or by the enabling technology such as touch gestures [134, 54] and computer vision-based hand gestures [123]. One good example of classifying hand gestures in human computer interaction is dividing them into the following five categories [57]:

Deictic: Deictic gestures involve pointing to establish the identity or spatial location of an object. Deictic gestures can be used with accompanied speech, e.g, by saying, "put that there" [11] or targeting a virtual auditory objects by pointing [73].

Manipulations: Manipulative gestures control an object with a direct relationship between the actual movements of the gesturing hand or arm with the object being manipulated. Manipulative gestures are used is various ways, e.g., with direct manipulative device such as a mouse to relocate or alter an object in a graphical user interface (GUI). Also gestures with other control devices or freehand gestures can be mapped to alter movement or rotation of virtual objects.

Semaphores: Semaphoric gestures are a set of distinct gestures that are used to communicate with a machine in a similar way as flags, lights, and arms are used in human interaction. They can be static poses such as the peace sign or dynamic hand movement drawing a circle in the air. Semaphoric gestures have a predefined and separate meaning and can be interpreted as commands. Semaphoric gestures can also refer to strokes or marks made with a mouse, stylus, or finger, such as strokes on a touch screen enabling eyes-free control of a music player [91].

Gesticulation: Gesticulations are hand movements that naturally accompany everyday speech and unlike semaphores do not have distinct vocabulary. They are not intended to be interpreted without the speech information that relates to them. Gesticulation can be used to communicate how something moves by saying, e.g, "the ball bounced like this" and use a hand gesture to show the trajectory of the ball.

Language gestures: Language gestures are a set of individual gestures the can be performed in series and are used to represent a language with a full grammar. An obvious example of language gestures is the sign language used especially by people with impaired hearing.

The gesture interaction in this thesis can be categorized as deictic or direct manipulation. Pointing gestures with auditory feedback has proven to be a working solution with hand gestures [23, 73], head gestures [15], and on a circular touch surface [138].

The enabling technologies used in this thesis work are the touch screen, devices with accelerometers, and camera-based tracking of a free hand. These technologies are used in three parallel interaction methods that rely on a circular interaction metaphor, in which the gesture is mapped directly to the position in a circular menu. A pointing gesture towards a menu item can be thought to be accompanied with a statement: "select this here".

Having one simple gesture has it advantages and disadvantages. One gesture is easy to learn and allows fast adoption of the new user interface. One gesture can also be used to control complex systems, just as in WIMP interaction (windows, icons, menus, pointer) a mouse is used to control computer GUIs. On the other hand, having multiple gestures can enable direct and more natural control of a system. For example, to start a song with a portable music player a "play" gesture could be performed without the need to find a play menu item or icon. However, the complex gesture vocabulary may be hard to learn and requires learning both on the part of the user as well as from the recognition algorithm [78, 59]. The set of suitable gestures may also vary between applications, and different people can prefer different gestures to perform the same task, thus multiplying the learning effort or making it harder to create universal gesture set [60].

The gesture interaction with auditory menus is discussed futher in Section 3.1. Next, the basic principles of human sound localization are discussed in the context of auditory menus and other spatial audio applications.

2.3 Sound localization and spatial sound

The sense of hearing differs from the senses of touch and vision, where the location information is mapped straight to the skin or the retina. The reason why humans and animals can localize sound sources well is because we have two ears located at opposite sides of the head. Sound localization is defined by Blauert [10] as follows: "Localization is the law or rule by which the location of an auditory event (e.g., its direction or distance) is related to a specific attribute or attributes of a sound event or of another event that is somehow correlated with the specific event". The locational information of a sound source has to be computed by analyzing the differences of two separate input signals. The auditory system has developed to be effective in the analysis of these spatial cues.

Interaural time difference (ITD) and interaural level difference (ILD) are the two main binaural cues that are used to estimate the horizontal location of the sound source [82]. The sounds arriving from either side of the head have different traveling times to the two ears, thus creating detectable difference. ITD presents the time difference of the wavefront arriving at both ears, caused by the fact that one ear is closer to the source than the other [82]. The phase difference of the arriving sounds can be identified from low frequency sounds, whose wavelength is longer than the space between the ears [82]. To give an approximation, the maximum delay for a 1 kHz sound is 0.65 ms for an average head. For sounds over 1.5 kHz, the wavelengths are smaller than the diameter of the head. Above this frequency the ITD does not help in localization [8].

The head serves also as an obstacle to the sounds, which causes sound level differences in the ears. This interaural level difference (ILD) is present in higher frequency sounds. The sound pressure level of high frequency sound is also encoded as electrical impulses that are used in the analysis done by the auditory system. ILD is the amplitude difference of the same wavefront, which is caused by the shadowing of the head [82]. The head is an obstruction to the sound, and for higher frequencies the in-

tensity of the sound is smaller for the ear that is farther away. Shadowing of the head starts to lose its effect for sound waves below 1.5 kHz, because they bend around the head and minimize the amplitude difference [8].

Another type of spatial information is the spectral cues produced mainly by the pinnae and also by the head and shoulders, which enable accurate location of sounds with complex spectral properties, even with one ear. Main purpose of the pinnae cues is thought to be the separation of front from back and as an elevation cue separation of up from down [8]. Information about sound elevation is obtained from frequency-dependent amplitude and time-delay differences caused by the pinnae [82] and pinnae can be thought as a linear filter whose transfer function changes depending on the distance and direction of the sound source [10]. The pinnae cavities change the travel time of the arriving wave front due to reflection, shadowing, dispersion, diffraction, interference, and resonance. Therefore some frequencies are attenuated and some amplified depending on direction of arrival, which can be used for extracting the location information. This effect is seen as distinct notches and bumps in the frequency responses measured from the ear canal [10].

The surrounding environment adds individual reflections and reverberation, which affects the perceived localization of the sound sources. Due to the precedence effect [128] (also known as the Haas effect [41] or law of the first wavefront [10]) humans can perceive the direction of the original sound source even if similar sounds arrive later from different directions. It allows us to localize sounds in the presence of reverberation, and the effect remains even if the latter sounds are louder than the original [41].

Sounds are naturally heard around us, and in order to reproduce spatial sound, ITD, ILD and other spectral cues have to be created. Sound reproduction can be typically divided into two approaches. The sound field to the listener's ears can be produced binaurally with headphones or with loudspeakers.

With loudspeakers, the created virtual sound source is already located away from the listener, and various methods can be used to position sound sources. Well known examples are stereo amplitude panning and vector base amplitude panning (VBAP) [93]. Both can be used to produce virtual sounds sources in the line or area between the loudspeakers. With surrounding loudspeakers, a virtual sound source is produced away from the listener, and the listening environment adds natural reflections and reverberation. Wave field synthesis [13] enables a more arbitrary sound

source positioning that is not dependent on the listener's position, but it requires large arrays of loudspeakers. Also recording and reproduction technique Ambisonics [72][38] can be used for producing a sound field and virtual sources.

Normal headphones or wearable speakers mounted on the user's shoulders [103] can be used while the user is moving and during daily activities. Wearable loudspeakers have the benefit of not covering the ears and the user can hear the surrounding environment, but also transparent headphones [76, 42] or bone conducting headphones [112] can be used. When using binaural reproduction with headphones (or only two loudspeakers), the binaural cues need to be reproduced artificially for positioning sound sources.

The ILD and ITD can be produced by appropriately delaying and attenuating the sound signals. The frequency-dependent attenuation caused by the shadowing of the head should be taken into account when designing a filter for ILD, for example, by approximating diffraction effects of a sphere [29]. The sound spatialization only with ILD and ILD can be sufficient for some applications, such as computer games. However, when using only ITD and ILD, the virtual sound source can be only located in a plane around the head, but a sense of source elevation is hard to achieve. Also, the virtual sound source produced with these interaural cues is prone to stay inside the listener's head [22] [8] [99].

The frequency-dependent amplitude and time-delay differences caused by the shoulders, the head and especially the shaping of the pinnae can be presented as head-related transfer functions (HRTF), which can be used for better localization and externalization of virtual sound sources. HRTF can be specified as the far-field frequency response for one ear, which is measured from a distinct point in free field to a distinct point in the ear canal [8]. With headphones, the HRTF functions can be used to produce binaural sound from a monoaural sound and positioned anywhere around the head. Adding artificial reverberation and reflections can help to externalize the virtual sound source even more, but at the same time localization accuracy may be decreased [6].

HRTFs vary a lot between individuals and HRTFs measured from other person's ears might not produce the same perceptual result [8]. Also general purpose HRTFs can be measured using a dummy head such as KEMAR, but using them can result front-back confusion or localization problems [8]. Unfortunately measuring individualized HRTFs is normally a

relatively laborious process that is done in an anechoic room and thus cannot be applied for everybody. However, recent research is trying to reduce the time and effort for designing individualized HRTFs, for example, it might be possible to take an image of a pinna and obtain the most important notch frequencies by analyzing the pinna anthropometry [96][108].

Although localization is not perfect with HRTFs, they still are a powerful tool and are often the principal spatial sound producing method with headphones. It is possible to create virtual sound sources that are perceived to be located around the listener. In aviation, dividing different communication signals into distinct spatial locations around the head helps the pilot to concentrate on the most important communication channel. Spatially positioned sounds are also efficient in guiding the pilot from the runway to the gate and they result in a faster response time when identifying a location of an external object that is causing a threat [7]. Spatial sound can produce a greater sense of immersion, discovery, and playfulness even in an auditory space with several sound sources [119]. Audio-conferencing also benefits from spatial sound adding social connectedness and group awareness [27]. However, with multiple audio streams, spatial sound can increase cognitive load if used improperly [120]. Furthermore, spatial sound is used to extend capabilities of small-screen devices and it may even provide better performance, for example, monitoring the progress of an event with spatial audio can be more effective than with a visual counterpart [124].

Virtual sound sources can be used as menu items, and sound positioning with ILD [91], ITD and ILD [138] and HRTFs [15] has been used with auditory menus. The correlation between the direction of reproduced sound and the gesture direction can help the user associate the sound with the specific menu item location [73]. Spatial sound can be used to separate each menu item better and to make them more distinguishable if, e.g., music is played at the same time. Spatial sound in auditory menus and gesture interaction are discussed in Chapter 5.

3. Related research

The previously published and related research work to this thesis are discussed in this chapter. First, the hand gesture interaction is reviewed by focusing on acceleration-based and recent camera-based eyes-free gesture interaction. Then, sound combined with gestures and spatial auditory menus are discussed with relevant examples from general auditory menu concepts, assistive technology, and vehicle systems. At the end of the chapter, interaction with performance spaces is reviewed with a few notable examples.

3.1 Auditory menus and gesture interaction

3.1.1 Hand gestures

Hand gesture detection can be roughly divided into two categories: 1) motion sensor-based approaches, where a sensor is attached to the hand being tracked, and 2) camera-based approaches, where gestures are recognized from images captured by a external camera. Various sensors can be used to detect tilting and other natural gestures [43]. However, this chapter concentrates mainly on accelerometer-based and camera-based interaction. Gestures can be also made on a surface, such as touch screens that are found in various everyday devices. Research on eyes-free touch gestures is discussed in the following sections, especially in Section 3.1.3.

An accelerometer-based approach can be used to detect gestures by holding a device with accelerometers in the hand. Nowadays, accelerometers are embedded in various devices, such as mobile phones, and they provide means for easy access to gesture control in everyday use. The Nintendo Wii remote control (Wiimote) has made the use of gestures popular, especially in gaming. Adding other sensors, such as gyroscopes, can make

the tracking more accurate and enable detection of more sophisticated gestures.

Different types of tilting interfaces using accelerometers have been presented mostly for visual displays, and many of them are methods for writing. TiltType [85] and Unigesture [104] include a writing interface where the tilt direction of the device can be used to specify letters. Similar systems have been proposed in mobile phones [131, 129], where the tilting direction and pressing the numeric keypad defines the output character. Tactile feedback has also been used in mobile phones with a one-dimensional tilt menu system to enable eyes-free use [79]. Tilting with a pen [115] has been used to to access a circular menu, where the pen tilt direction is used to select visual menu items. One of the first systems using tilting interaction was presented in 1996 by Rekimoto [97]. He utilized FASTRACK position and an orientation sensor and applied tilting and two-button-device to browse menus on a visual display.

The interaction with wrist rotations of a horizontally held arm has been studied with a multipart mobile device consisting of a SHAKE sensor pack attached to the hand as a wristwatch and a Nokia N95 for visual feedback [26]. Wrist rotations have proved to be quite an accurate and feasible control method, but simultaneous walking can make interaction harder [26]. Furthermore, studies about controlling applications with wrist tilts include interaction with a handheld device, remotely interacting with a screen [95, 3], and audio-only music browsing [113] using a Wiimote to navigate in a large music collection.

Free-hand gestures with camera-based tracking can offer an interaction method in which no devices need to be attached to the user [58], ideally allowing control without any preparation from the user [123]. Gestures enable touchless interfaces that allow operation from a distance and can be used when there is risk of contamination, e.g., in hospitals [122]. The use of hand gestures can bring safety benefits when using a vehicle's secondary controls (e.g., radio or heating) by reducing the need to reach out for objects inside the vehicle and maximizing eyes-on-the-road and hands-on-the-wheel times [88]. Just like Wiimote popularized acceleration-based gestures tracking in gaming, The Microsoft's Kinect [62] is used for free-hand and body gesture recognition for interaction in games.

Free-hand gestures are applied both in desktop and mobile contexts. SmartCanvas [74] is an intelligent desktop system allowing free-hand drawing with two cameras. It uses pie-shaped menus and a finger rota-

tion for menu selection by taking advantage of the finger's proprioception. The menu is activated by extending the thumb, browsed by rotating the index finger (i.e., changing its roll), and a selection is made by maintaining the orientation of the finger for few seconds. Mo et al. [74] argue that moving a finger to the location of a menu item for a selection would be inconvenient, because it requires the user to coordinate the fingertip motion on the desk with the motion of the pointer on screen. This could be avoided, e.g., by accessing the menu items by moving the fingertip to the relative direction from a defined center point.

A visual and markerless detection of fingertips on a mobile phone can enable gesture detection while on the move [4]. Menu selection can be implemented using a camera-detected fingertip moving a cursor or overlaying the actual hand on the image in real time [4]. Imaginary Interfaces [40] is an example of a mobile interface where a camera is attached to the user's chest (e.g. necklace). The system allows performing spatial interaction gestures with empty hands and without visual feedback. One hand is used to give a reference with an L-shaped gesture creating coordinates for the second hand. The study of Imaginary Interfaces suggests that free form interaction and a button selection are possible with screenless interaction. An example of an eyes-free gestural interaction is Virtual shelves [66]. It enables pointing and selecting objects that are located in virtual positions in front of a user. Virtual shelves shows that eyes-free pointing gestures can be used in spatial interfaces.

Free-hand gestures can be combined with sound in various ways. Examples include simply controlling music playback [67], creating music itself [114, 58], and using sound to deliver information [44]. A predefined set of free-hand gestures to control music playback has been developed with comparative evaluation by Löcken et al. [67]. They used camera-based tracking to identify dynamic and static hand gestures. They argue that predefined gestures must be carefully designed to be usable instead of relying on the user's or the designer's intuition. However, gestures still need to be learned before use and different sets of gestures are needed for different usage scenarios. Camera-based tracking can be used to detect a user's hands in location (x, y), posture and angle of rotation to create interactive computer music performances [114]. This kind of system allows many degrees of freedom to control music and express emotions through gestures and music. In this approach, the hands themselves are the instrument [114], but hand gestures can also be used to control virtual instruments

such as a guitar [58]. Gestures can also be utilized in sonification, which in contrast to speech interfaces, uses non-verbal sounds to present information. Gesture Desk [44] uses arm and hand gestures to extend desktop interaction with the mouse and keyboard. Gestures are tracked in three dimensions with a camera through a glass table. The system is used to find patterns or regularities in data by interactively browsing through an auditory map.

3.1.2 Auditory menu concepts

Pirhonen et al. [91] tested a prototype of an eyes-free touch interface for a simple music player, in which music playing was controlled with finger sweeps on the screen. The finger sweeps from left to right, top to bottom, and vice versa were used to control the volume and change the music track. Tapping of the screen was used to start and stop the track. Their study pointed out that immediate audio feedback is vital for user confidence, and the interface proved to be effective in eyes-free situations.

Audio is utilized in eyes-free user interfaces, and circular auditory menus in particular have been extensively studied. Horizontal circular auditory menus are favored because localization of horizontally positioned sounds is more accurate than vertically positioned ones. Savidis et al. [101, 25] used the concept of auditory windows where a subset of four sound objects was simultaneously played in a spatially larger area, while others were pressed closer together. They used pointing interaction with a data glove, a head tracker, and voice recognition to control a modifiable circular auditory environment reproduced with headphones. Kobayashi and Schmandt [63] used an egocentric circular interface to access temporal audio data, such as simultaneous audio recordings. Work by Friedlander et al. [35, 34] showed that circular "bullseye" menus can be effective with audio-only feedback. They used simple beeps without spatial sound to indicate menu items and a stylus or mouse as a pointing device. Brewster et al. [15] used a directional head-nodding interface to study four simultaneous auditory menu items located around the user. They showed that head gestures are a successful interaction technique with egocentric sounds. Other notable examples applying circular auditory menu are the Nomadic Radio [103] and a calendar application by Walker et al. [125]. The study of Marentakis and Brewster [73] on audio target acquisition in the horizontal plane concluded that pointing interaction with spatial

sound is successful with mobile users. They also suggested that audio elements with feedback from egocentric auditory displays may produce efficient designs. Visual circular menus also outperform standard pull down menus [20] and are widely used in the user interfaces of computer programs.

Speaking the menu items one by one is a traditional way of menu navigation, which is largely used in interactive voice response (IVR) systems in telecommunications, but because of slowness and lack of user control it is frustrating in active use [137]. Zhao et al. [138] emphasized the importance of instant reactivity to user input. Their usability studies with touch input and a circular touchpad showed that an auditory menu can outperform a typical visual menu used in iPod-like devices. Their study did not compare visual and auditory interfaces when the input gestures are the same, although visual circular menus have been reported to improve both seek-time and error-rates over pull down menus [20]. However, the Earpod [138] interface combined many useful features from previous research and introduced important ideas such as 1) direct reactivity to touch input that gives control to the user without waiting periods, 2) interruptibility of the audio, where only one sound is played at a time, but its playing can be interrupted if the user chooses to continue browsing, and 3) menu items which can be accessed directly without browsing through all items.

The Foogue concept [28] is an example of eyes-free interface with gesture input that does not require visual attention. Foogue can be used to control a mobile device in two modes: menu mode and listening mode. Menu mode is for browsing and controlling a file system that is presented with spatial sound in front of the user. In the listening mode, music, phone calls, and auditory notifications can be heard simultaneously and positioned around the head of the user. If fully implemented, Foogue might allow eyes-free control of a mobile phone, and complementing it with a visual interface is possible.

3.1.3 Auditory menus in assistive technology

Touch screens in mobile phones, home appliances, and public facilities can create difficulties for visually impaired users. One of the main problems is that the visually impaired users cannot efficiently locate the graphical user interface elements on a flat surface [117]. The voice-over screen reader of Macintosh computers (OSX) and on the iPhone (iOS) can make

touch screen interfaces accessible to visually impaired users. Still, touch screens are primarily designed for persons with normal vision, and the use of voice-over might not be the most efficient solution. The interfaces can be designed also in terms of audio and implementing completely different interfaces for the sighted and visually impaired engaging different sensory modalities is also justified [133]. Using modality-independent interaction patterns to design user interfaces and leaving the implementation open is possible [31, 32]. Thus the strengths of each modality can be used in the interface implementations that can be totally different depending on the modality used.

The sonic grid [48] was developed to help visually impaired people correlate the movement of a pointing device with the corresponding location on the screen and to perceive the spatial layout of a GUI. The sonic grid gives sound feedback on the position (or coordinates) in the screen with non-speech audio cues. The horizonal position is encoded by stereo panning, and pitch is changed for vertical positioning. The sonic grid was reported to be effective, but also requiring a long learning time before it could be efficiently used.

In a gesture-based text entry method for touch screen devices called NaviTouch [39], all letters are accessed through vowels. The user first slides his finger vertically to find vowels that are read out loud. After hearing any of the vowels (e.g. A), the user can slide his finger horizontally to find consonants that follow that particular vowel in the alphabets (e.g. B or C). The user makes one L-shaped gesture for each successful consonant selection.

Kane et al. [55] used a similar L-shaped touch-gesture to browse music tracks. In the reported experiment, ten album names were placed vertically in a list. Each item on the list could be listened to one at a time. The user first found the desired album with a vertical finger swipe and continued the finger movement to the right to hear the track names. A second finger tap was used to select items. Although only one continuous touch-gesture can be used to access songs, it does not solve the problem when the music library holds hundreds of albums.

No-Look notes, introduced by Bonner et al. [12] used multi-touch text entry with the aid of a circular pie menu, which was shown to be much better than using a QWERTY button arrangement with the iPhone's built in voice-over. Bonner et al. suggested that a successful eyes-free text entry system needs to incorporate 1) a robust entry technique, 2) a familiar

layout, and 3) painless exploration.

Kane et al. [56] also studied how gestures differ between sighted and blind people to understand better how to build touch screen interfaces that work equally well for blind and sighted people. Blind people may prefer different gestures and they also may perform them differently than sighted people. Kane et al. reached the same conclusion as Bonner et al. that using robust gestures that reduce demand for location accuracy is important as is the use of familiar spatial layouts.

Text entry can also be implemented using a different touch screen gesture for each character [116, 71]. Tinwala and MacKenzie [116] used gestures that resemble letters as input and auditory and tactile feedback to guide eyes-free entry. Letters were entered one at a time, and word-level error recognition with a dictionary was used to improve accuracy. Tinwala and MacKenzie suggested that changing the speech feedback from the character-level to the word-level did speed up writing and lessen user frustration. The method was evaluated to be reasonably fast and accurate. In the auditory menu of Tinwala and MacKenzie, the word suggestions were spoken with 0.6 s breaks, and the user could pick the correct one. Due to good error correction, most of the words suggested to the users were either in first or second position.

Bezel menus can be also used for eyes-free menu navigation and writing [49]. Bezel menus are used by crossing the boundary of a touch screen with an inward swiping gesture. When writing, the boundary crossing position (e.g., corners or sides) first defines a set of four letters, and then a swipe direction selects one of the letters from a pie menu. The study by Jain and Balakrishnan [49] suggest that accurate eyes-free interaction is achievable with a layout of 32 (8 times 4) menu items.

Braille text is also used with touch screen interaction [50, 100, 33]. Braille touch [100, 33] allows touch-typing without sliding the fingers on a touch screen. It is used with six fingers so that each touching finger represent a Braille dot and an auditory confirmation is used for each typed character. The V-braille system [50] was designed for deaf-blind users for reading braille. The phone vibration is used to represent Braille dots when a finger is placed on one of the six regions on a touch screen.

3.1.4 Auditory menus in cars

Auditory displays can improve the usability of the system when eyes-free operation is necessary, and they are useful while driving and when visual attention should be focused on the road.

Sodnick et al. [107] studied in-vehicle interaction with auditory and visual menus. They used a scrolling wheel and two buttons attached to the steering wheel to move a selecting bar in a visual list-style menu or to rotate a circular auditory menu. The design did not allow direct access to all menu items on one menu level. Instead, in the auditory menu, all auditory menu items were rotated around the head and the one in the front was the one to be selected. They also evaluated auditory menus with the simultaneous sound sources versus one sound source at a time. Although the auditory interfaces were not faster, they were found to be effective to use, improved driving performance in short tasks, and lowered overall work load.

Jeon et al. [51] carried out dual task study where the participants simultaneously played a ball catching game comparable to driving and navigating a song list. They added auditory menu cues to a visual menu. Both performance in the game and the menu search time were better with the auditory cues than with no sound. Jeon et al. suggest that auditory cues can help drivers to maintain their attention on the road more effectively than visual-only menus.

3.2 Interaction with performance spaces

This section reviews few notable and relevant examples how interaction with performance spaces is used in theatre, opera, and orchestra settings. Technology can be used to augment a dance performance or a theater stage with interactive content that responds to movement and gesture in credible, aesthetic, and expressive ways [109].

An early example of audiovisual interaction with narrative environments is the play It/I [90], where one of the main characters was played by a computer and included virtual stages. It was one of the first attempts to create a computer controlled, interactive, and story-based environment, where real actors or spectators could interact with a virtual actor. Cameras were used to track the actors, enabling interaction with the virtual

actor. The virtual actor would follow the script of the play and act accordingly by recognizing gestures of the actors belonging to certain scene. The abstract virtual actor was a screen projection, who communicated with the user by combining synthesized sound, images, movies, and lighting effects.

In the opera Jew of Malta [65], the arts organization and design studio Art+Com did a sophisticated combination of a projected stage and projections on costumes that were accurately mapped to the actor's movement. This large production included virtual architecture that was projected on several movable projection screens. Dynamic costumes were also projected and changed in real-time on the white clothes of the moving actors. The actors could interact with the virtual stage, for example, by moving and gesturing with an extender arm. The main goal of the interaction with the virtual stage was to support the narrative, where the main character is losing his ability to control the world.

An example of interactive music and sound is the MIT media lab's Brain opera [81, 83], which was a touring interactive production used for performances and public installation. The audience could interact and generate music through gestures and touching the futuristic instruments. The interaction sensors included electric field sensors, touch pads, springs, touch screens, various tactile interfaces, inertial sensors, and optical trackers. Sensors were used to manipulate sounds and samples of a user's voice. The installation explored the concept of responsive environments, where ubiquitous technology in the surrounding environment detects physical activity or motion and creates a multimedia response.

Sound can be also used to augment a performance hall, especially to enhance the acoustics for special purposes. With electro-acoustic enhancement, the rehearsal room acoustics can be changed to match the performance hall as well as possible. This is advantageous because symphony orchestras do not have the possibility to always rehearse in the concert halls where they perform. Lokki et al. introduced a system that would allow a symphony orchestra to play in an electronically augmented rehearsal hall [68]. The system included a reverberation enhancement system with microphones and loudspeakers. It addressed the problem of uncontrolled feedback loops by using time-varying algorithms for the artificial reverberation. With time-variance, the loop transfer function of the system varies continuously, which prevents the self-generating peaks and allows higher gains without coloration and instability. The system

was evaluated with professional musicians in a small multipurpose performance hall [69] by using the system to give them a feeling of interaction in a much lager space while playing. The musicians found that the artificial reverberation is quite well suited for an orchestra practice hall and no coloration from the time-varying algorithm was perceived.

4. Interaction with virtual spaces

Publications I and II are examples of how eyes-free interaction with the surrounding space itself can be implemented. The next section describes a multi-wall reverberation system, which is used to modify the size and shape of an acoustic space. In Section 4.2, the same reverberation system is used in an actual opera and performance hall. Section 4.2 also reviews how gestural interaction was used in a live performance to interact with a virtual stage.

4.1 Reverberation system

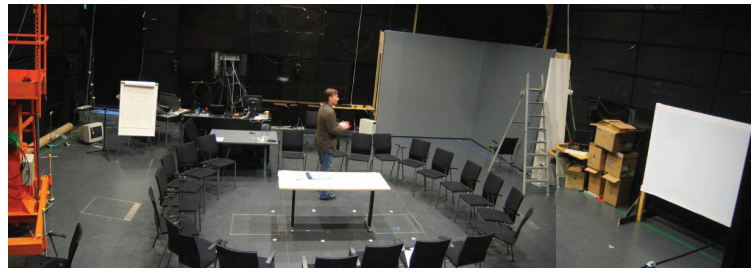

Figure 4.1. Test subject doing the subjective experiment. The subjects could freely move inside the area marked out by chairs.

Here, a room containing multiple reverberation enhancement systems (RES) is presented. The whole setup consists of four individual systems, one on each wall of an acoustically dry room. The ”reverberation time of each wall”, meaning the reverberation time (RT) in each direction, is controlled individually, thus it is hypothesized that the size and shape of the perceived auditory space may be controlled. The provided auditory space has no physical counterpart, it is not known, and thus it is interesting to see how people will perceive the auditory space.

The test environment (shown in Figure 4.1) is a multipurpose facility designed to be a lecture hall or performance theatre, but also a laboratory for experimental tests of virtual reality technology and applications. Its volume is about 790 m^3 ($12m \times 11m \times 6m$). The walls and ceiling have been constructed to be very absorbing while the floor is hard concrete. The reverberation time is short compared to the volume, and the computed absorption coefficients are high.

Each RES consists of a microphone in the center of a wall, a time-variant reverberator, and six loudspeakers. The loudspeakers (Genelec 1029A) are mounted so that three are at ear level, two are elevated about 40 degrees, and one is almost directly above (see Figure 4.2). The loudspeakers are approximately on the surface of a hemisphere, but to have them exactly at an equal distance from the center of the room, they were virtually positioned by using delays. In addition, the loudspeaker gains are adjusted to obtain an equal sound pressure level from each loudspeaker at the center of the room.

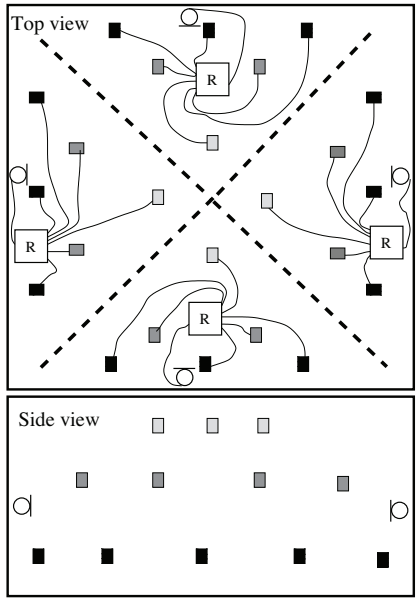

Figure 4.2. The loudspeakers are positioned approximately on the surface of a hemisphere so that 12 of them are at ear level, 8 are about 40 degrees elevated and 4 are almost on the ceiling. The signal routing is drawn with lines from the microphones mounted in the center of each wall to the reverberators (R) and to the loudspeakers.

The computer controlling all four RES simultaneously has an RME DIGI9652 sound card that offers three ADAT digital I/O ports. All the four micro-

phones are connected to a 8-channel AD converter which feeds the controlling PC via one ADAT interface. The output from the sound card is routed to three 8-channel DA converters that directly feed the active loudspeakers.

The whole system is implemented with the Pure Data graphical programming language [92, 87]. It is simple to use and enables low latency with the applied sound card and ASIO drivers in Windows XP. With the Pure Data audio buffer size of 10 ms, the latency is about 17.5 ms, measured by looping the output directly to input.

4.1.1 Evaluation of the virtual acoustic environments

The task of the subjects was to test eight different virtual acoustic environments. The test cases do not correspond to any real acoustic space, but they were obtained just by applying different reverberation times in different directions. Another choice to control the acoustic shape would have been to apply different a gain to each wall, but this was not studied. The intended shapes of the virtual acoustic environments were a square, a rectangular, and an open wall, as depicted in Figure 4.3. To create such shapes the reverberation times for each direction (i.e., for each wall) were defined as listed in Table 4.1.

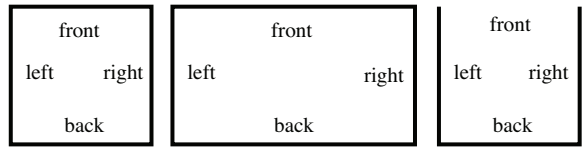

Figure 4.3. The tested virtual acoustic spaces as seen from above. See Table 4.1 for parameters applied for each wall.

In total, ten subjects (7 males, 3 females) completed the test and drew the perceived auditory shapes. All subjects except one had academic background and the average age of the subjects were 30. The subjects did not know any details of the purpose of the study nor the setup. They entered the room (not blindfolded) and were given short instructions to test eight different cases. They were free to make any sound to evaluate the shape and size of the auditory space. Subjects were advised to talk, shout, or sing in all directions and to clap their hands or to make any noise. The area in which they could test the system was a space of about 4 times 4 meters in the middle of the room (see Figure 4.1).

Case	Shape	RT at walls [s]				Extra
		right	back	left	front	
2	square	0.0	0.0	0.0	0.0	-
3	square	1.5	1.5	1.5	1.5	-
6	square	3.0	3.0	3.0	3.0	-
8	square	3.0	3.0	3.0	3.0	40ms
1	square	3.0	3.0	3.0	3.0	70ms
4	rectangle	2.5	0.5	2.5	0.5	-
7	rectangle	2.5	0.5	2.5	0.5	1.5dB
5	open wall	2.0	2.0	2.0	0.0	-

Table 4.1. Parameters of the eight test cases. In cases 1 and 8 extra delay was applied to each wall. In case 7, the gains of the rear and front walls were about 1.5 dB higher than in case 4.

In the beginning, subjects were instructed to test all cases, before they started the real evaluation. They could freely decide the order in which the spaces were evaluated. In addition, subjects could listen to each space as many times as they wanted. Results were given by the subject drawing on the answer sheet the perceived shape and size of the auditory space. The test was completed when all eight cases were evaluated and drawings for all eight cases were ready.

In all, 80 answer sheets (ten drawings for each case) were collected, and they were scanned into electronic form. Then each case was treated separately, and the drawings were overlaid (see the results in Figures 4.4–4.7). In addition, quantitative measures were taken from each drawing by measuring the area and aspect ratio of the drawn shapes. The average and standard deviation of these measures are presented in Table 4.2.

First, the drawings in the case when all RESs were off (case 2) is seen in the top part of Figure 4.4. A few subjects drew the lines representing the walls exactly over the walls of the floor plan, but most subjects perceived the room to be smaller that in reality. This was expected, because the room is almost semi-anechoic, and it really sounds smaller than it is. The second result is seen at the bottom of the same figure (case 3, RT = 1.5 s at each wall). It seems that most subjects perceived case 3 as a square room, but a few subjects also perceived a rectangular room. It might be that the rectangular form of the answer sheet encouraged subjects to draw more rectangular shapes, because the mean of the aspect ratios is close to that of an A4 sheet, see Table 4.2. The estimations of the sizes vary from

Case	Area	Std dev. (area)	Aspect ratio	Std dev. (asp. ratio)
Floor plan	42	-	1.05	-
A4 sheet	624	-	1.41	-
Case 2	14	16	1.20	0.5
Case 3	74	59	1.37	0.5
Case 6	236	144	1.21	0.4
Case 8	278	142	1.33	0.3
Case 1	310	143	1.33	0.2
Case 4	171	121	2.32	1.7
Case 7	144	83	2.82	2.0
Case 5	163	123	1.63	1.1

Table 4.2. Areas and aspect ratios of drawings in each of the 8 cases. The values are the averages of the ten subjects.

smaller to larger than the real physical space, the mean being bigger than the floor plan of the physical room.

With a long reverberation time of RT = 3.0 s at each wall (cases 1, 6, and 8), the room was perceived to be large, as seen in Figure 4.5. However, the standard deviation between drawn areas is also large (see Table 4.2). The biggest drawings are more rectangular than square; perhaps again the form of the A4 sheet led subjects to draw rectangles. There is also one 3-D drawing in case 8, where the subject claimed to perceive a high space. Cases 8 and 1 included extra delays of 40 ms and 70 ms, respectively. This does not seem to change the perception of shape, but they had an effect on the mean of the areas (see Table 4.2).

Almost all subjects perceived the rectangular shape that was being created in cases 4 and 7, see Figure 4.6. One subject even left the front and rear walls open by drawing only the left and the right walls. When an extra gain of 1.5 dB was applied to the front and the rear walls, the drawings are more consistent with each other. It should be noted that although the aspect ratios in cases 4 and 7 are larger than in other cases, the standard deviations are also larger.

Finally, the open wall (case 5) results are shown in Figure 4.7. Only one subject left the front wall undrawn. However, all front wall drawings are close to the floor-plan front wall, while other walls were perceived to be more distant.

The test cases were reported to sound natural by all subjects, and they could imagine themselves being in a real space. A few subjects found minor artifacts such as unnatural echos with impulsive sounds or “something unnatural” at the end of the reverberation tail. The echoes were heard in cases when extra delay was added before the reverberation algorithm. Thus, extra delays should not be applied although they seemed to help to make the perceived auditory space slightly larger.

The results indicate that the implemented system can produce natural sounding virtual auditory environments and that applying different reverberation times in different directions can modify the size and shape of the perceived space.

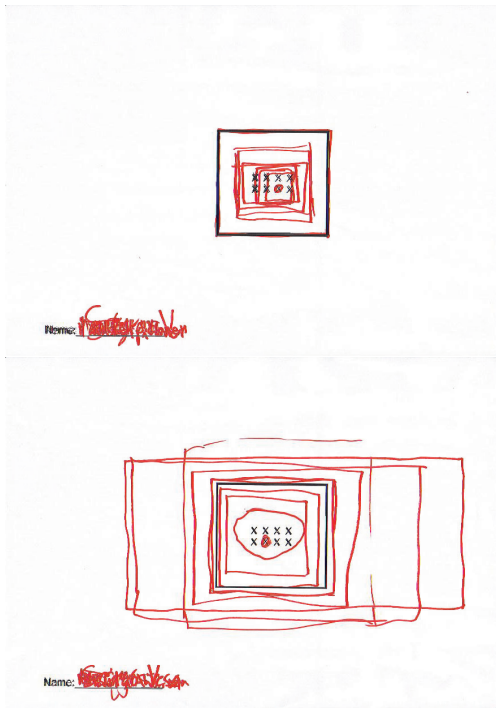

Figure 4.4. Drawings by the ten subjects of the perceived auditory space. Case 2 is above and the case 3 is below. The crosses in these and in the following figures are information about the test case and do not relate to the perceived size or shape of the spaces.

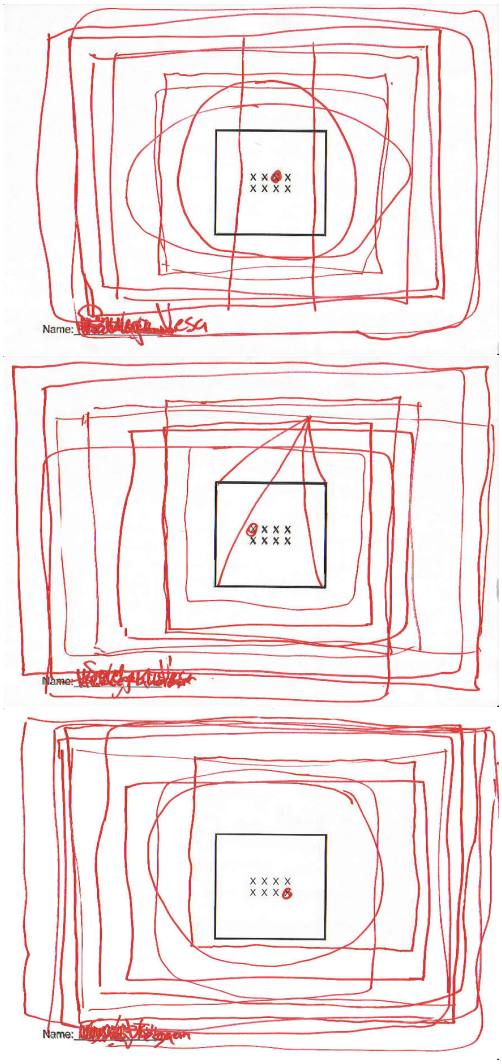

Figure 4.5. Drawings by the ten subjects of the perceived auditory space: case 6 (top), case 8 (middle), and case 1 (bottom).

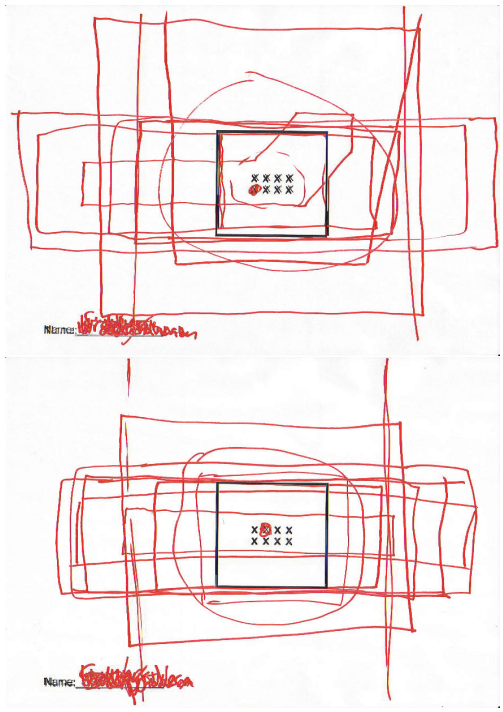

Figure 4.6. Drawings by the ten subjects of the perceived auditory space: case 4 (top) and case 7 (bottom).

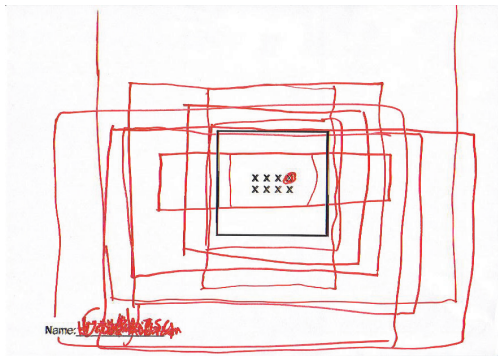

Figure 4.7. Drawings by the ten subjects of the perceived auditory space in test case 5.

4.2 Application: Virtual opera

Figure 4.8. Digitally augmented stage with actors and musicians with instruments.

This section reviews an experimental opera production, where digitally augmented content was used interactively during the performance. Projected graphics and spatial sounds were designed to support the story (see Figure 4.8).

In a traditional stage performance, the effects are carefully timed by a technician to the performers' actions, creating an illusion of interaction. In this opera performers were given more freedom with a aim to study wether the performance benefits from real interaction with the virtual stage. The actors of the opera concentrate intensly on their performance and developing an easy, and eyes-free control of the interactive stage was important.

The production was made in a collaboration between Helsinki University of Technology, Helsinki Institute for Information Technology, Theater Academy, and Sibelius Academy. The writing of the libretto, the music composition, and the virtual stage development started simultaneously and progressed in parallel. This way, the music could affect the story and the virtual effects, and vice versa. For example, the sounds imitating a light switch were written in the musical score where the actor cuts off electricity from Europe. The actual performance included a conductor, two percussionists, two pianists with grand pianos, and two singers (Alice and Vorotov). The lights, the virtual stage, and the subtitles employed

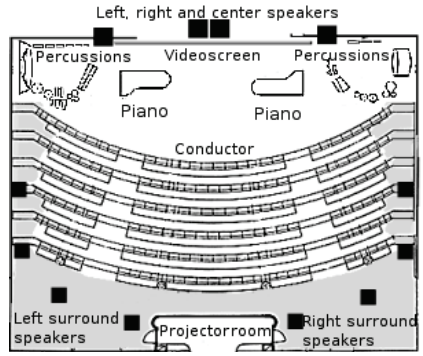

Figure 4.9. Layout of the performance hall in the Ateneum art museum, Helsinki.

three extra persons during the performances. The physical setting of the performance hall is illustrated in Figure 4.9

The libretto of the opera is loosely based on Anton Chekhov's novel Expensive Lessons, where a rich young man suffers from inability to speak French in the 17th century. The libretto was rewritten to reflect the foreign policies of Putin's Russia, where a young Mafia gangster lacks social skills, preventing him to master international crimes. The opera tells the story of two characters with different backgrounds. Vorotov is the juvenile son of a mafioso, whose illicit business is restricted to operate only in Moscow. Vorotov hires Alice to teach him social skills, but falls fatally in love with her.

Various scenes worked as narrative elements in the play. They visualized the environment or revealed the thoughts of the characters (see Figure 4.10). This way complex interpretations of the story could be presented to the audience. The visual expression was also linked to the musical interpretation.

One main visual concept was a 3-D octagon (more precisely a rhombicuboctahedron), which could hold different images on its 18 faces. The textures could be changed to the each side of the octagon generating a platform that was used in several scenes of the opera. In the story, the octagon can be seen to have many metaphors. Vorotov can be seen to spin the octagon to show off the wealth placed outside of the octagon. He is in control of the world and the octagon behaves as he wants. On the other hand, the female character Alice is prisoned inside the octagon and subjected to inhumane treatment by persons controlling the octagon. Its development began from an idea of a traditional rotating theater stage

which is used to change the scene in the opera. The random spinning of the octagon can be also seen as reminiscent of a Russian roulette or wheel of fortune. The metaphor is made even stronger by using audio effects of wheel clicks and a spinning revolver cylinder. The spinning sound around the audience was created by playing clicks in one 5.1 loudspeaker group at a time. The sound spun in the direction defined by the visual octagon. The original idea was to use VBAP [93] to control the exact position of the sound. This was used in the research facility, but the loudspeaker system in the performance hall did not allow separate control of the surround speakers. Control of the roulette was handled with the Wiimote. First, a way to spin the roulette in any direction using acceleration was developed, but horizontal and vertical spinning was more useful for the performers. The speed of the gesture was analyzed using the acceleration sensors of the Wiimote, and the octagon was spun accordingly, giving it a certain speed and a random stop position. Inside the octagon the spinning was made more arbitrary by using the pitch and roll of the Wiimote.

The opera performance consisted of several scenes that combined traditional and new interactive methods of control. Each scene had individual functions that could be controlled either from a computer or by using the Wiimote. In the first scene, the Russian flag was molded by the music recorded by the ceiling microphone, which was analyzed, and the transients generated the fluctuating bubbling of the flag with a physical wave propagation model (see Figure 4.10). A virtual Earth was used to show the view of the world that Vorotov controlled, stating his power and that of Russias. The illicitly traded goods are animated on the surface of the globe. The world could be spun with the Wiimote, giving it more speed with a hand gesture, or rotating it more precisely with arrow buttons. The buttons were also used to zoom and trigger animated effects. The use of Wiimote had to made robust against errors (e.g., requiring specific button combinations to trigger effects), since in this scene the performers were arguing over the device.

The sound reproduction system was used to modify the acoustics of the performance hall. A time-variant reverberation enhancement system implemented with a feedback delay network was used to modify the reverberation in the hall, allowing higher gains without causing instability of the system [68]. The reverberation system described in Section 4.1 was adapted for the performance hall and was used with the loudspeaker setup shown in Figure 4.9. Although the pre-placed loudspeakers limited

the implementation, it was possible to add more reverberation from the direction of the left surround speakers, right surround speakers or the front speakers. The physical reverberation of the performance hall affected the perceived auditory environment by the audience and it was not compensated in any way. Due to the artistic nature of the performance, a physically more accurate reverberation was not pursued after. The goal was to create enhanced acoustics for each scene that would support the narrative and create a whole together with instrumental music. The enhanced acoustics of each scene were experimented with the composer, and predefined settings were used during the performance. The reverberation was changed to suit the location or mood of the story, and in some places to create a dialog between the instruments and the sound production system.

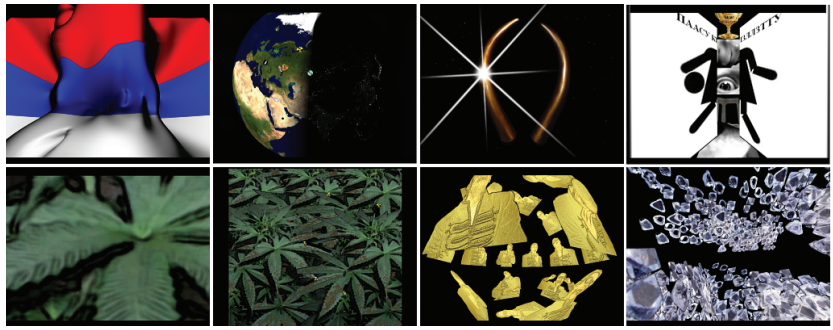

Figure 4.10. Screen-shots from the interactive virtual stage environment used as a narrative element in the play. The virtual stage was projected on a large screen behind the stage.

4.3 Discussion and lessons learned

The reverberation system described in Section 4.1 was frequently used to produce support for speech in the multipurpose lecture room and later in the opera described in Section 4.2. The multipurpose room is acoustically so dry that the voice of a lecturer sometimes gets exhausted due to the lack of acoustic support from the walls. When the reverberation system is on, and the reverberation time is between 1.0 s and 1.5 s at each wall, the room sounds more natural and many visitors don't even notice that the acoustics is enhanced electronically.

The reverberation system was used in a contemporary opera in which

acoustics of the performance space changes between acts, and where other virtual acoustic effects was added to the performance. In the opera performance, the reverberation system was controlled only by a technician, but using gestures to change the reverberation parameters for an on-demand space and voice alteration was experimented with during rehearsals.

The performer-based interaction is useful when user controlled media is directly mapped to gestures and when detailed nuances of movement are hard for a technician controlling the media to follow. However, using technology just to trigger effects within a completely rehearsed performance lies the danger of art becoming a show of technological prowess. In fact, opera and other performing arts with more room for improvisation can make the most of the media controlled by performers. The motivation behind the opera was to study the benefits of the interactive approach and also pure curiosity for virtual opera technology.

Technology can also bring new problems. In the opera, the performers were in control of the virtual stage. The use of technology can raise the mental load of the performers, since they have to concentrate on new things. The performers would have required more rehearsal time to get better acquainted with the technology. The rehearsal time was limited to about 45 hours, and the technology was not available in all rehearsals. In an ideal situation, technology adopts to the performers' needs and not vice versa.

The gestures and how the Wiimote should be used were discussed with the performers during the rehearsals. The first rehearsals revealed that the gestures and button combinations should be simple. The time allotted to the rehearsals defined which actions performers had time to adopt and which were controlled from a computer. The performers did not want too much responsibility, being afraid of making mistakes or loosing concentration. Therefore, a technician controlled the largest changes in the virtual stage, such as changing of the scenes.

Smaller sensors than the Wiimote would be useful. Although the Wiimote was part of the act as a sceptre of power, the performers complained that in some scenes having a lot of action, the Wiimote was inconvenient. A better solution would be to integrate acceleration sensors and buttons in the performer's sleeve, where they seamlessly integrate with the clothing. Sensors such as accelerometers, joint-angle sensors, heartbeat sensors, temperature sensors, light sensors, and image sensors can be connected with a wireless network for real-time gesture and movement monitoring

[84]. This way the hands of the performer are free for expression, but their motion can still be accurately tracked.

The gesture interaction in the opera with Wiimote motivated the research of eyes-free interaction methods and auditory menus described in the next chapter, Chapter 5.

5. Interaction with auditory menus

The following sections summarize Publications III, IV, V, and VI, which introduce advanced circular auditory menus and describe three parallel control methods for their using. In Publications III, IV, and V, new control methods and auditory menu properties are tested with user experiments. Publication VI presents a mobile application that integrates auditory and visual menus.

5.1 Circular control method

The introduced interfaces are controlled with circular gestures, that can either be made with the wrist by holding a device in the hand (see Section 5.3), on a touch screen (see Sections 5.4 and 5.5), or free-hand interaction in the air (see Section 5.5). These three parallel interaction methods rely on a circular interaction metaphor that enables natural interaction with circular menus. Ideally the user can choose any of the three control methods to access the same circular menu.

One emphasis in the interaction design is that the same input can be used to control both visual and auditory menus. Therefore, a circular menu is used where all menu items can be accessed with the same effort, and the item position can be learned so that they are always found in the same direction. A circular menu is accessed by moving the hand around a predefined center point or by moving the hand directly towards the menu item.

Furthermore, a music player application that enables efficient eyes-free control was built as a proof of concept to show that complex tasks can be performed with circular auditory menus (see Section 5.6). A similar design can be used to control other functions of emerging modern devices.

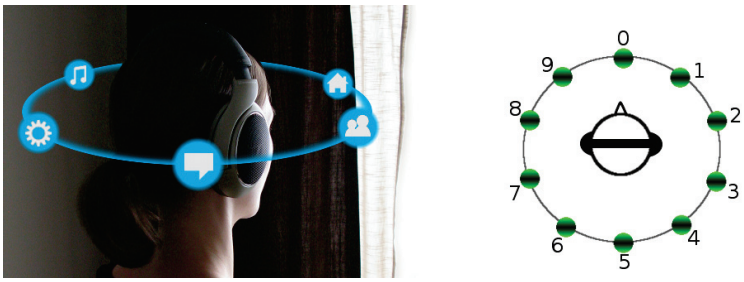

Figure 5.1. General concept of the auditory menu visualized. The auditory menu items are positioned on a virtual circle around the head of the user.

5.2 Auditory menu

The work described in the Publications III, IV, V, and VI use the same circular auditory menu concept, which is improved in each publication. This section gives an overview of the used circular auditory menu.

The general concept of the used auditory menu is depicted in Figure 5.1. The circular auditory menu is reproduced with spatial sound using either headphones or loudspeakers. The auditory menu items are positioned around the user's head with spatial sound. The menu is egocentric, meaning that the menu is positioned relative to the users body or head. The menu items are words or sentences spoken with synthesized speech, and other sound effects can be mixed with them.

The used circular auditory menu is based on previous research described in Chapter 3 and especially that of Brewster et al. [15] and Zhao et al. [138]. The key element is the use of interruptible audio and an immediate response to the user's input with an auditory display. The spoken menu items are played one by one while browsing a menu, and the user has the ability to jump to the next item thus stopping the playback of the previous one. With slower motion, the user can hear all menu items one by one. Thus, the user is in control and he or she can adjust browsing speed according to his or her own abilities.

As mentioned before, Bonner et al. [12] suggested that a successful eyes-free text entry system should include a robust entry technique, a familiar layout, and painless exploration. The same applies to browsing the introduced eyes-free auditory menus. The menu can be browsed with a simple and robust circular motion and a selection can be made when the desired menu item is heard. Furthermore, when the concept of the circular menu is familiar, the users immediately know how the menu is laid out. In ad-

dition, placing the items in alphabetical order can be used to ease the use of the menus.

Advancing in the menu hierarchy is done by selecting a menu item from the circle and reversing by making a selection in the center of the circle. This design was chosen for consistency. Always, having a "back" menu item present would occupy space from the circle. The center of the circle is easy to find during eyes-free use, because the circular gesture goes around it and the user is constantly aware of its position. When the center of the circle is reached, the name of the current menu level is read out loud and mixed with a short "bubble pop"-like auditory icon indicating that the center is now active. This makes it possible for the user to always query the location in the menu structure as Kane et al. [55] also suggested. After a short delay, the name of the higher menu level is read out loud and the user can thus traverse in the menu structure.

When browsing faster, the user hears only the beginning of the sounds. Because the short sounds (or phonemes) represent the first letter(s) of the names, they help the user keep track of the position in a large menu. This feature has been recently evaluated as beneficial and has been suggested to be named "spindex" [53, 51]. In Publications III, IV and VI, the spindexes are automatically generated when the user browses the menu. This is achieved through the auditory menu's instant reactivity to users' gestures. By slowing down the browsing speed, the user can adjust the length of the spindex thus enabling an efficient search method for menu items, starting with the same letter, letters or even word.

Spatial sound helps to distinguish sounds coming from different directions. The localization is good on a horizontal plane and the reproduced sound directions can help the user to associate the sound to the specific menu item location [73]. Spatial sound can also give better understanding of the shape of a menu and improve the performance. Proper design can also improve the performance as the user gets familiar with the spatial menu item configuration. Furthermore, each menu item is heard from a different spatial direction, making it easier to distinguish them when browsing with increased speed.

The binaural implementation for headphone reproduction applies HRTFs, which enable a more realistic reproduction of the spatial sound localization cues. The sounds are also processed with a simple reverberation algorithm, which helps in the externalization of auditory menu items. In addition, two artificial early reflections are created by attenuating and

delaying the direct sound and by reproducing them from different horizontal positions. The HRTF data were measured with the method designed by Pulkki et al. [94], where a loudspeaker was rotated around the subject with continuous movement in an anechoic room, and responses were measured with a swept-sine technique [30]. This process produced HRTFs every 6° in azimuth and every 15° in elevation in elevation angles between -30° and 45°. However, the horizontal circular menus only need earlevel HRTFs. The implementation of HRTF filtering uses minimum-phase HRTFs and a separate ITD model. The ITD is computed with a spherical head model, and minimum-phase HRTFs are modeled with 30-tap long FIR filters, as presented by Savioja et al. [102]. The interpolation between measured positions is done separately with fractional delays for the ITDs and linearly for finite impulse response (FIR) filter coefficients.

Normally the sound sources produced with HRTFs can cause localization problems such as front-back confusions, where a sound source in the front of the head is perceived in the back or vice versa. This was not observed to affect the use of the auditory menu, although this was not explicitly studied in this thesis. Furthermore, the localization of the auditory menu item is made easier by always reproducing the sound from the direction of pointing. The user is already expecting to hear the sound from the back when he or she is pointing backwards, thus reducing the effect of localization problems such as front-back confusion.

The combination of gestural interaction and human spatial hearing also enables more complex 3-D menu structures. Menu items can be positioned on several elevated horizontal levels, enabling simultaneous access to item groups. In Publication III, a simple improvement was used in the user experiment, where a "back" menu item was positioned directly above making it fast to access. Positioning in 3D can support a larger number of items, and in some cases easier selection. Figure 5.2 shows a dial menu where special menu items are grouped in the upper part of a sphere and numbers are on the horizontal plane. In this approach, the 'sight' is restricted to stay on one of the menu item groups.

It is important to give feedback to the user when a selection is made. There are many suggested non-speech feedback sounds, e.g., auditory icons [37], earcons [9, 16], and spearcons [127, 126]. The implementation presented in this thesis, uses a fast replay of the selected menu item mixed with a short auditory icon. A short clink sound is played immediately after the selection, followed by the fast replay of the selected menu

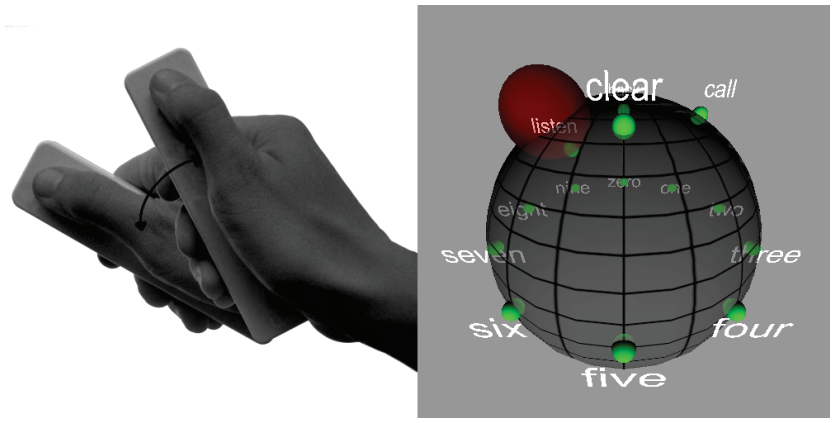

Figure 5.2. Visualization of a 3-D menu layout. The selection between upper special items and lower numbers can be done by adjusting the tilting angle of the control device.

item. The playback time of the sound is shortened considerably, but the user can still easily recognize the content. The clink sound further clarifies that the selection was made. The changed pitch also indicates a feedback sound, not another menu item. In this way, the user gets immediate feedback and can easily double-check whether a correct selection was made. During the research for this thesis, simultaneous speech sounds were found to create confusion for inexperienced users of auditory displays, and the participants mixed the confirmation of the selection and menu items. In the study in Publication V, the menu item sounds were delayed after the selection sounds. This is supported also by previous literature [107], which concludes that multiple simultaneous sounds should be avoided in particular during higher cognitive workload with reduced concentration.

To enhance the selection accuracy of a menu with several items, a dynamically adjusted target sector, where the item is active (played), can be applied. As visualized in Figure 5.3 (left), if none of the items is active, the menu items have a target area with same size. When a menu item is active, its target area expands in both directions reaching a 1.9-times larger target area. The value of 1.9 was chosen to leave a big enough target area for the neighboring menu items, because they shrink, making room for the expanding sector. This is done to facilitate easier browsing and selection by reducing undesired jumping between tightly packed menu items. The positions of the menu items can also be adjusted dynamically which is explained in Section 5.4.

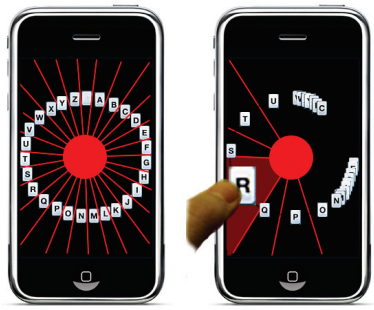

Figure 5.3. Menu items are defined as dynamically changing sectors on a screen. The visualization is exaggerated and artificially created. It is not needed in the eyes-free auditory menu.

Auditory menus also need to have working controls which are easy and safe in eyes-free use. In Publication VI, an advanced volume control is used. When the volume menu item is selected, the user can adjust the volume with a circular slider by using again the circular motion and making a selection for accepting the change (see, Figure 5.18c). The volume slider was designed so that the volume cannot be accidentally turned to the maximum level. A user who is unfamiliar with the menu might start the exploration from any part of the screen. The volume adjustment is done relative to the starting position and not from a fixed position on the screen. Furthermore, when the volume is lowered, jumping accidentally to maximum volume is not possible. Instead, the end of the volume slider follows the gesture until it stops. A similar control menu can be used for other continuous control or searching.

5.3 Accelerometer-based interaction

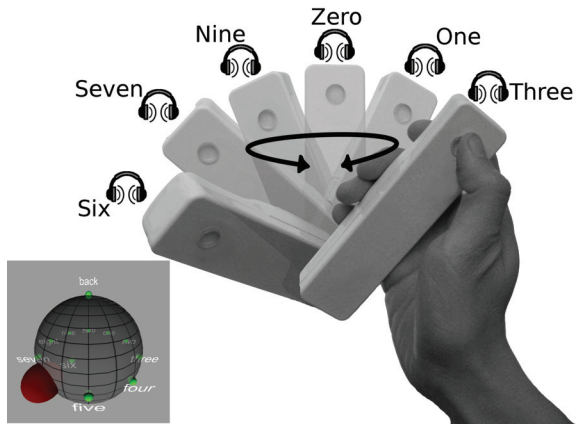

Figure 5.4. Gesture interface utilizing a mobile device mockup. Auditory menu items can be accessed by pointing or tilting the device in the desired direction. The design can be further improved by placing the button on the top and shaping the device for better grip.

Publication III introduced a novel control method which uses accelerometers along three axes to find the orientation of the device relative to the direction of Earth's gravity, as depicted in Figure 5.4. The Nintendo Wiimote was used to the user experiment for record the acceleration for gesture recognition and its button data for selection.

The mapping of acceleration data from the Cartesian coordinated (x, y, z) to spherical coordinates (azimuth and elevation) enables an easy way to define position on the surface of a sphere. The acceleration data from one axis varies between 1 and -1, being 1 when the axis is downwards, 0 when lateral and -1 when facing up. If the device is exposed to higher accelerations, the input data is clamped between -1 and 1 to keep the values on the surface of a sphere. Normally when using accelerometers for exact positioning, a problem occurs when the device is rotated perpendicular to Earth's gravitational field and acceleration sensors cannot detect this motion. This limitation is avoided when the device is used as shown in Figure 5.5.

In the case of the auditory menu, the use of the control device as if it were a joystick is convenient. This is done by tilting the device slightly and rotating it 360° with a gentle wrist gesture, as shown in Figure 5.6. The tilt angle needed to access the menu items can be only 5°, allowing small wrist movements and preventing any tedious turning of the wrist.

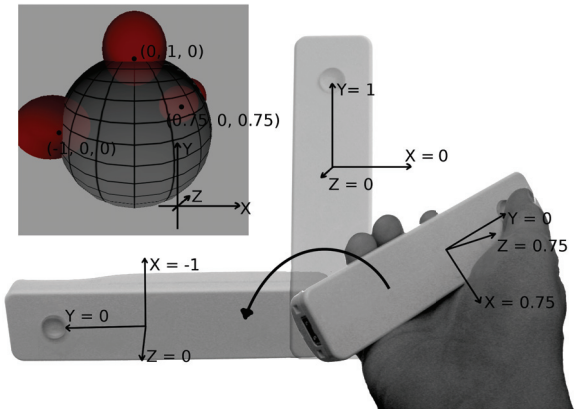

Figure 5.5. Pointing in any direction in the user-centered 3-D space is possible by using data from only three accelerometers. A wrist movement gesture guides the device to tilt correctly, and the right way of pointing is learned fast. The red "sights" on the surface of the sphere visualizes the pointing direction. The coordinates of the "sights" are converted from acceleration sensor data affected by gravity.

5.3.1 Evaluation

User experiments, with 11 normal-hearing subjects, were conducted to determine the accuracy and speed of the implemented system. All subjects were right-handed males with academic background and age varying from 24 to 38 years. Subjects did not have previous experience about audio interfaces. Three simple tasks to test the usability of the gestural interaction method together with auditory menus were chosen. Within-subjects design was chosen for reducing number of the needed test subjects and reducing errors associated with individual differences.

First the subjects were shortly advised how to handle the Wii Remote to successfully tilt and point to desired direction. Then the subjects practiced the use of each menu layout for about five minutes by writing 30 characters. Visualization of menus and pointing was shown with a desktop monitor only in the beginning of the training. Then subjects used the system with audio cues only. The menu layout in the training session was the same as in the experiments, as it was preferable that subjects knew the menu layout beforehand.

Figure 5.7 shows a visualization of the auditory menu used in the three (T1, T2, T3) user experiment tasks:

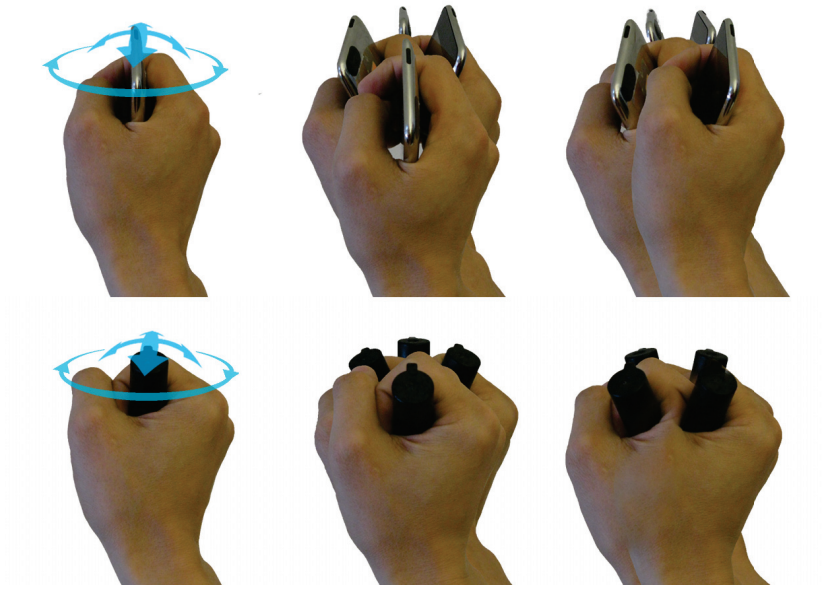

Figure 5.6. The wrist gestures are performed by slightly tilting the device towards the desired menu item, as if the device were a joystick in the air. The browsing can be continued with a circular gesture. The figures illustrate the upright position (left), tilting towards the cardinal points (center), and the half-cardinal points (right). A device shaped like a tube would fit the hand better and is more ergonomic to use (below).

- *Seated and selecting ten random 10-digit numbers with a menu consisting of numbers. (T1)*

- *Seated and selecting a random series of letters or real words consisting of 9 to 12 letters, each with a menu consisting of 26 letters and a "spacebar"-item ordered alphabetically from a to z. (T2)*

- *Walking and selecting ten random 10-digit numbers with a menu consisting of numbers. (T3)*

- *Seated and calling 10 randomly chosen persons existing in the phone book with a menu of multiple levels representing a phonebook. (T4)*

The fourth task (T4) tested a more realistic use scenario of a mobile phone. The tasks (numbers, words, and names) were displayed on a desktop monitor, but no visual feedback about the success of the selection was given. A preliminary mobile user experiment (T3) was conducted to the test if presented interaction method suffers when being used in a mobile

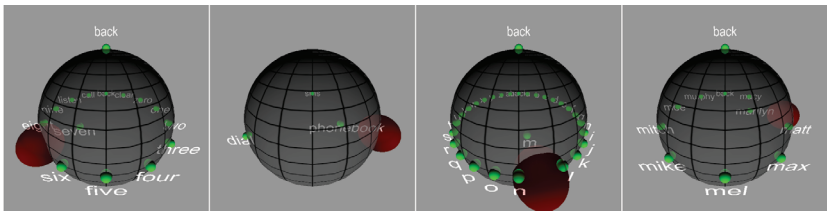

Figure 5.7. Visualization of the auditory menu levels. A simple mobile phone layout included the possibility to dial numbers, send an SMS, and browse names from an alphabetical list. The green menu items represent the spoken auditory menu items and their spatial positioning in a circular auditory menu. The bigger red ball shows the tilting direction and the currently active auditory menu item.

scenario. This time five subjects performed only the number selecting task, but during the test the subjects walked in a figure-of-eight around two chairs placed about 3 meters apart, the method used in [125, 15].

Results and analysis

The total collected data consisted of selection times and correctness of 1100 selected numbers (T1), 1155 written letters (T2), and various data from the fourth task (T4). The percentage of correct numbers and letters was 99.4% and 95.1%, respectively. Since the distribution of raw selection times was positively skewed, the median time of each block (10 numbers or 9–12 letters) was only used to compute the mean of medians. The means of selection times were 2.13 s for a number and 3.83 s for a letter (see Figure 5.8). The success rate in the calling task (T4) was 100% since the subjects had the possibility to go back and fix any error in item selection. If only those cases when the users got it right the first time are considered, the success rate drops to 92.7%. The median (and mean) time to make one call took 13.76 s (15.20 s).

Despite walking in the preliminary mobile user experiment (T3), the selection accuracy was 98.3% (two subjects made no errors at all). The mean of the median selection times for a number was 2.28 s. As seen in Figure 5.8, there is no significant difference in selection times of immobile (T1) and mobile (T3) use, confirmed by the analysis of variances ($F = 1.96$, $p = .163$).

All subjects easily learnt how to operate the Wiimote and to use it to browse menus. Observations, recorded data, and discussion with the sub-

jects revealed that they could immediately find the desired menu items without the need to browse through all items. In the number writing task (T1), the first gesture frequently caught the right menu item, but in the case of alphabets, the first gesture gave the general direction and the right menu item was found after browsing through a few neighboring elements. Entering letters was harder because of the larger menu. The subjects also mentioned that remembering the order of the alphabets takes longer than numbers.

The selection speed of the presented system is comparable to the earPod [138], which has been shown to be faster than the traditional visual iPod menu. With the earPod, the selection time of one menu item from among eight options was 1.9 s and accuracy 94.2%. In this experiment, similar results (2.13 s) were achieved with higher accuracy (99.4%), despite a larger menu (ten items). The presented method also outperforms the earlier eyes-free interaction methods in the number of items present in the menu. The possible reason for this might be the accurate pointing implemented with wrist rotations. It seems that the kinetic memory of the human hand is very accurate and giving feedback with spatial sound is intuitive. Another study on wrist rotations suggests that a 9° resolution could be achieved in target acquisition [26]. Additionally, the mapping from wrist movements (input) to spatial sound (output) seems to work well. Thanks to static menus, the test subjects easily learned how to directly access menu items, despite there being more than ten items. It is possible that earlier studies with circular auditory menus have suffered from simultaneously audible menu items and low resolution interaction methods, such as head nods [15] or a small touch-input device [138].

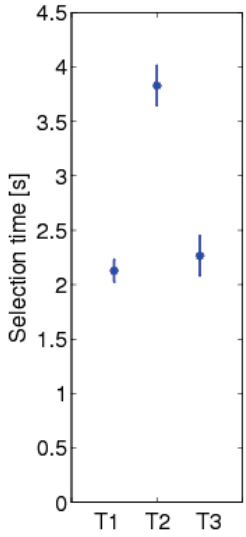

Figure 5.8. Means and 95% confidence intervals for medians of selection times. T1 means selecting one number, T2 selecting one letter, and T3 is the same as T1, but in a mobile scenario.

5.4 Touch screen interaction

Publication IV extended the control interaction to touch screen use and also tested the accelerometer-based input with an actual mobile device. The main contribution of the Publication IV is the dynamic menu item spreading, which enables fast and efficient browsing of long lists in circular auditory menus.

A touch-surface (or a screen) can be used to access a circular auditory menu (see Figure 5.9), and sectors extending from the center of the surface represent the menu items, as shown in Figure 5.3. The user can access any item directly by placing a finger on the surface and can continue browsing with a circular finger sweep. Removing the finger from the surface makes a selection. The center of the touch-surface is a safe area from where the finger can be lifted without making a selection. No selection is made if the finger accidentally slides off the touch screen area, which might happen especially during eyes-free use.

5.4.1 Evaluation

Nine participants completed an experiment where a touch screen and gesture interaction were used to access large menus. All participants were

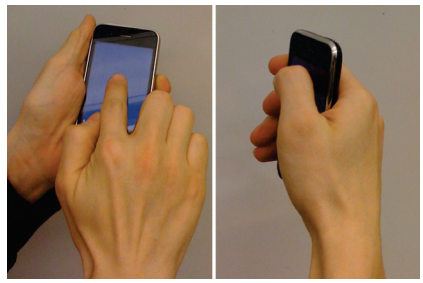

Figure 5.9. Two interaction methods were implemented as prototype applications for the iPhone. In the touch interface, the auditory menu was accessed by touching the screen and gliding the finger on it. A selection was made by removing the finger (left). In the gesture interface the auditory menu was browsed by tilting and rotating the device while touching the screen with the thumb. The selection was made by releasing the thumb from the touch screen. No visualization of the auditory menu was shown in the device.

males between 23 and 43 years old. The participants volunteered for the experiment, and they had no previous experience with the interaction methods and auditory menus used. The goal of the experiment was to study the selection speed and accuracy of auditory menus with a large number (>100) of items. Two alternative auditory menu layouts containing a contact list of 156 (26 times 6) names was used. As will be explained later, one menu layout was a more traditional auditory menu with two layers and the other used a novel approach to fit all 156 names in one menu level. Earlier studies with different interaction methods had suggested that egocentric auditory menus could contain at most five [70], eight [15], or twelve [138] menu items in usable scenarios. This is probably due to limitations in the interaction devices, browsing methods or simultaneously played sounds. With the presented auditory menu layout, the number of names displayed to the user could be dramatically increased when compared to the number of names in Slide Rule [55].

The auditory menus were reproduced with headphones, and the names to be chosen with different interaction methods were listed on a large projection screen. An iPhone was used as the test device. The connection between the iPhone and the laptop was implemented by using the Open Sound Control (OSC) protocol and a modified version of the free Mrmr software [75] installed on the iPhone. The auditory menu was implemented using Pure Data (PD) [92, 87], which received the raw control information from the iPhone.

Within-subjects design was chosen for reducing number of the needed test subjects and reducing errors associated with individual differences.

The actual experiment consisted of four tasks using auditory menus and one visual task giving reference time and accuracy. The five tested methods were:

- *Reference* (Ref): the normal contact list of the iPhone without any auditory feedback.

- *Touch screen one-layer* (T_1L): the eyes-free touch screen input with 3-D audio output. All names were directly accessible in one menu layer(see Figure 5.10).

- *Touch screen two-layers* (T_2L): the eyes-free touch screen input with 3-D audio output. The subject initially selected the first letter and then the name from a submenu (see Figure 5.11).

- *Gesture-based one-layer* (G_1L): the eyes-free gesture input with 3-D audio output. All names were directly accessible in one menu layer (see Figure 5.10).

- *Gesture-based two-layers* (G_2L): the eyes-free gesture input with 3-D audio output. The subject initially selected the first letter of the name and then the name from a submenu (see Figure 5.11).

In the user experiment, the participants walked around four aligned chairs while choosing names from a large menu. Before every task the participants were allowed to briefly test the interaction method using a set of 5 names, which remained the same for all methods. The practice time was restricted to 5 minutes. In the actual task, 11 slides containing 5 names were used. The next slide was revealed right after the last name of the previous slide was completed. The participants were instructed to carry on to the next name, even if they made a mistake. The names were Finnish first and last names.

The order of the tested methods was randomized between the participants to ensure proper control groups. In all methods, the participants browsed the same contact list containing 156 names, 6 names for each of the 26 alphabets from a to z.

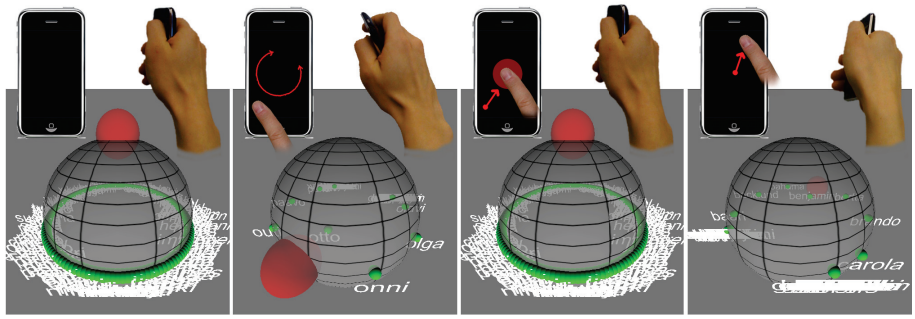

Figure 5.10. Method for browsing large eyes-free auditory menus. It combines qualities of absolute positioning of menu items and browsing easiness of smaller menus. The browsing method can handle hundreds of items that can still be accessed rapidly. Absolute positioning of menu items is used when the control device is pointed upwards or when the touchscreen is not touched (left). The desired menu items starting with same alphabet are always found from the same direction. When any of the items is active the other menu items are spread to have much wider spacing, and browsing can be continued to either direction. The menu items can be laid out to absolute positions by moving the finger to the center of the screen or by pointing the device up. The browsing method can handle hundreds of items that can still be accessed fast.

Menu with two layers: The general layout of the two-layer menu is visualized in Figure 5.11. The first menu level consists of 26 letters from A to Z, which are always found in the same locations and are placed in alphabetical order. By selecting a letter, the user can advance to the second layer of the menu consisting of names.

Menu with hundreds of items in one layer: The novel one-layer menu layout can handle a very large number of items, as shown in Figure 5.10. The applied browsing method combines the benefit of a-priori known item positions in a static menu with large menus. In this approach, when none of the items is selected the menu items are in their absolute positions in alphabetical order. For example, all the names starting with the letter A are placed in alphabetical order in the sector that occupies the letter A in the menu shown in Figure 5.10 (left). Thus, the user can point to or touch the desired position and hear one of the names starting with that letter.

When the user targets a particular item, its neighboring items are spread around evenly with a spacing of 40° , and items farther away are grouped together (see Figure 5.10, right). If the desired menu item is not found directly, the user can continue browsing items with a rotating hand gesture or a circular finger sweep. The next item is always found 40° forward and

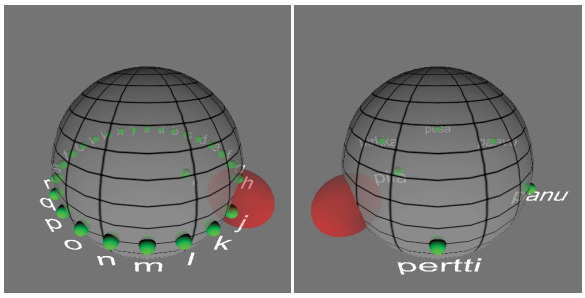

Figure 5.11. The two-layer menu. The alphabets are always found from the same position in the first menu layer. The second layer holds 6 names in alphabetical order and they are spread evenly.

the previous one 40° backwards, respectively. This spreading can also be seen as a different implementation of the fish-eye distortion concept [111, 36].

Results and analysis

The correctness of selected names is presented in Table 5.1. The means of the median selection times are listed in Table 5.1 and shown in Figure 5.12. A two-way analysis of variances yielded a main effect for both methods, $F(4, 405) = 154.77$, $p < .001$, and the participants $F(8, 405) = 4.60$, $p < .001$.

The advanced menu item spreading in the one-layer menus (T_1L, G_1L) was proven to be effective in the experiment. There are many benefits in this approach. First of all, there is only one menu level and selection needs to be made only once when searching. This can reduce the possibility of an error and increase the selection speed. Also, the distance between menu items is always the same regardless of the number of items, which facilitates the browsing. Furthermore, this novel menu layout enables fast transition from one part of the list to another.

After the study, the dynamic layout was further enhanced to achieve faster and better usability. One improvement is included in the application described in Section 5.6, in which the starting place is always defined

Table 5.1. Results of the user experiment.

	Ref	T_1L	T_2L	G_1L	G_2L
Correct selections [%]	97.0	96.4	95.8	88.6	87.6
Selection times [s]	3.43	7.02	7.98	9.65	10.76

to be the first name in alphabetical order. When the user points the device or touches the screen, for example on the letter D, he or she would always know that names starting with D will be heard when browsing clockwise, and when counter clockwise browsing the last name starting with C is found.

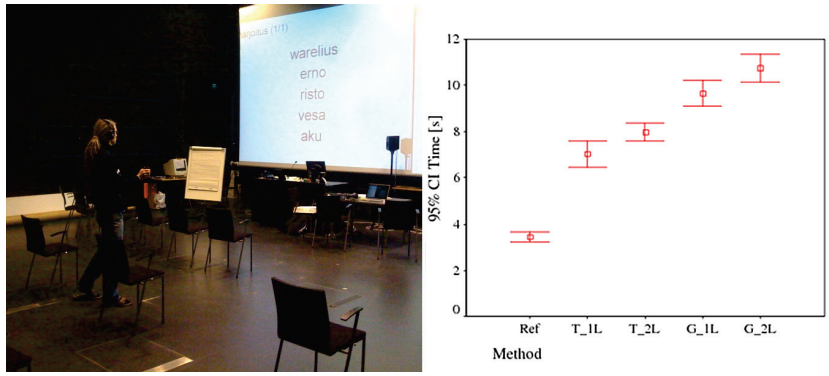

Figure 5.12. Left: The mobile user experiment. Right: Means and 95% confidence intervals of selection times of all interaction methods.

5.5 Free-hand interaction

Publication V extended the control interaction to free-hand gestures with camera-based tracking and also tested the touch screen-based input with a larger screen size. The main result of Publication V is that the free-hand gesture interaction is possibly faster with an auditory menu than with a visual one.

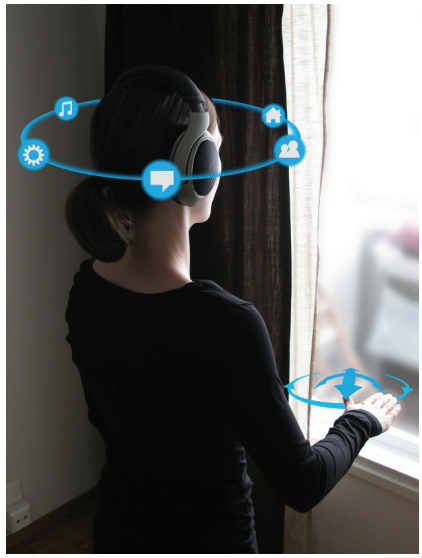

Figure 5.13. Concept of free-hand controlled circular auditory interface

Gesture-based systems should be accessible without extensive training [123]. Complex gestures, such as drawing a triangle in the air or different hand poses the need to be memorized, but here the idea is to use simple gestures without need for extensive training. Figure 5.13 illustrates the general concept of how a circular gesture can be used to match the auditory menu items reproduced with headphones. The angle around the fixed center point is used to define the angle for accessing menu items in the circular menu. Menu items can be either browsed with a circular hand gesture or by moving directly to the desired menu item.

A wide variety of mechanisms can track a user's gestures, but this experiment used a Kinect sensor [62] because of its capability for depth recognition. The Kinect was positioned 150 cm above the surface of the table and its cameras faced the surface directly. The Kinect was used to track the position and distance (x, y, z) of the fingertips. The distance information from the Kinect was used for the selection gesture. Figure 5.14 shows

the camera views of the Kinect and the result of a simple hand-tracking algorithm. The selection gesture should be as easy and natural as possible, and selection by pushing toward an item has been used in various systems. The implemented selection gesture can be thought of as pushing and releasing a virtual button. A simple algorithm was developed to indicate selection by monitoring the distance of the fingertips from the camera. The algorithm identifies that a selection is made when the distance of a tracked point grows continuously (pushing), reached a local maximum (button pushed to the bottom), and the distance starts to decrease (button released). One of the reasons for choosing this selecting gesture was its naturalness and that works with one finger or a full hand.

In a preliminary study, different approaches for the gesture interface were tested, and a vertical movement of the arm was found to cause fatigue, even if the interface is used for short periods of time. This is also known as the ”gorilla-arm effect”, which occurs when interacting with vertical touch screens. Making the gestures horizontally causes less fatigue especially when the gestures are done low and there is no need to keep the hand raised. The horizontal gesture fits well in many use scenarios, such as in a car, on a surface display, and on a desktop environment.

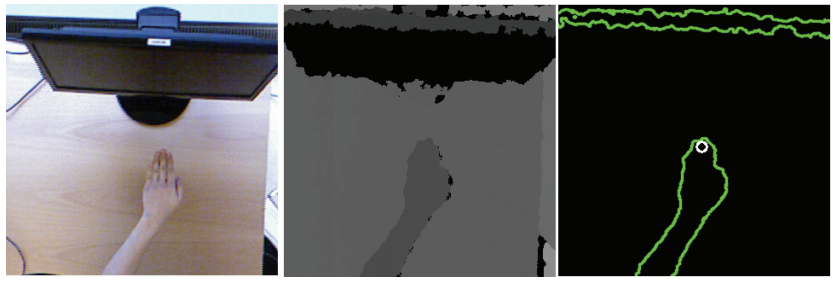

Figure 5.14. Simple hand tracking algorithm implemented for the user study by using Kinect. From left to right: 1) RGB image of the hand, 2) depth image of Kinect, and 3) detected contours and a white circle marking the fingertips and the point being tracked.

5.5.1 Evaluation

Fifteen subjects (4 females and 11 males) participated to the user study. All participants, except one, had an academic background and their ages varied from 27 to 46. The actual experiment consisted of six tasks using auditory and visual menus measuring time and accuracy. With all methods, participants wrote 10-digits long random numbers, which did

not contain the same digit consecutively. The numbers to be selected were shown on the LCD display approximately at eye level, and the participants completed 100 digits using each interaction method.

Before each task, the participants could practice the interaction method. The practice was similar to the actual experiment and participants typed at least two 10 digit numbers until they felt confident enough of the interaction method to proceed the actual task. This was done to ensure that, 1) there would not be a long pause between the practice and the task, 2) the participant is familiar with the upcoming task, and 3) the participant is confident enough with the interaction method. However, longer practice time would be needed to minimize the effect of learning during the actual task.

One iPad and one Kinect sensor were used as inputs. The gestures used for control were identical in the visual and the auditory menus. Visual and auditory menus were not used at the same time. Within-subjects design was chosen for reducing number of the needed test subjects and reducing errors associated with individual differences. The experiment was a simple factorial design, in which six different interaction methods were tested. The methods were:

- *Touch screen and visual circular menu* (TC_V): the touch screen input with a visual display and a circular menu.

- *Touch screen and auditory circular menu* (TC_A): the eyes-free touch screen input with a spatial auditory display and a circular menu.

- *Gesture and visual circular menu* (GC_V): the gesture input with a visual display and a circular menu.

- *Gesture and auditory circular menu* (GC_A): the eyes-free gesture input with a spatial auditory display and a circular menu.

- *Touch screen and visual numpad menu* (TN_V): the touch screen input with a visual display and a numeric keypad, or numpad, menu.

- *Touch screen and auditory numpad menu* (TN_A): the eyes-free touch screen input with a mono auditory display and a numpad menu.

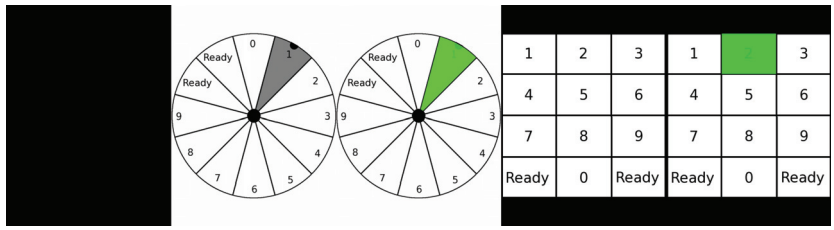

Figure 5.15. The iPad screen was used as a visual display and an input device for the touch screen interaction. When using an auditory display, the iPad screen was black (left). Circular and numpad style menus were used in the experiment.

When an auditory display was used, the iPad screen was black, and in the visual menu the iPad screen displayed either a circular or numpad menu, as shown from left to right in Figure 5.15.

The primary research question was how an auditory display compares with the visual counterpart when the input method remains the same. The secondary research question was to compare the speed and accuracy of the six methods. The usability of the circular touch screen interface has been already tested in the previous study described in Section 5.4, and now the free-hand gestures (GC) and a larger touch screen (TC) were tested when using the same circular menu.

The numpad menu (TN) was designed to be the reference interface, since the most participants would be very familiar with the layout. It gave a credible reference for the speed and accuracy for all participants. However, the use of the numpad with free-hand interaction was evaluated to be slow in the preliminary testing, and it was not included as one of the interaction methods. The user study was designed to test a scenario without an additional cognitive task, but still requiring the participant to move his or her eyes between the input device and the stimulus.

The order of the tested methods was randomized between the participants to ensure proper control groups. To randomize the task order, TC, GC, and TN were paired and the order varied between participants. Further more, every second participant started either with auditory display (A) or visual display (V).

Results and analysis

Because the distribution of all raw selection times was positively skewed, the median selection times of each digit were compared with a non-parametric one-way analysis of variance. First, the differences in selection times be-

Table 5.2. Mean and the median times for one selection with all interaction methods.

	TC_V	TC_A	GC_V	GC_A	TN_V	TN_A
Mean [s]	1.15	1.63	2.40	2.28	0.82	1.45
Median [s]	0.92	1.39	2.19	2.04	0.61	1.22

Table 5.3. Number of correct selections by each participant. The percentages of each participant (far right) and of each interaction method (bottom) are given.

	TC_V	TC_A	GC_V	GC_A	TN_V	TN_A	Total [%]
1	100	94	93	97	99	98	**96.8**
2	96	99	100	97	99	100	**98.5**
3	95	100	100	96	98	100	**98.2**
4	97	96	93	98	100	95	**96.5**
5	97	95	87	95	94	100	**94.7**
6	100	100	90	98	99	99	**97.7**
7	99	98	98	100	98	96	**98.2**
8	97	100	99	99	100	100	**99.2**
9	98	100	99	98	100	98	**98.8**
10	100	100	97	100	100	99	**99.3**
11	100	99	98	96	100	98	**98.5**
12	99	99	98	97	100	99	**98.7**
13	100	99	98	96	100	96	**98.2**
14	96	97	97	92	98	93	**95.5**
15	100	100	95	96	100	100	**98.5**
ALL [%]	**98.3**	**98.4**	**96.1**	**97.0**	**99.0**	**98.1**	

tween individual participants was found to be large, as shown in Figure 5.16. However, all participants performed consistently with all interaction methods. A Kruskall-Wallis one-way analysis of variance shows significant differences between the rank means (χ^2= 418.73, $p < .001$).

The differences between rank means of tasks were also analyzed with the Kruskal-Wallis procedure. The rank means differ significantly (χ^2= 3683.7, $p < .001$). Post-hoc analysis using Tukey's least significant difference criterion ($p = .05$) for the six conditions showed this difference to exist between all cases, except between GC_V and GC_A. The means and the medians for each interaction method are shown in Table 5.2 and median times in Figure 5.17. The number of correct selections by each participant is presented in Table 5.3.

The movement during the selection gesture and distance from the center were measured with gesture interaction. There were large differences in the performance between the participants. The median angular movement during the selection gesture across all participants was 2.86 (GC_V)

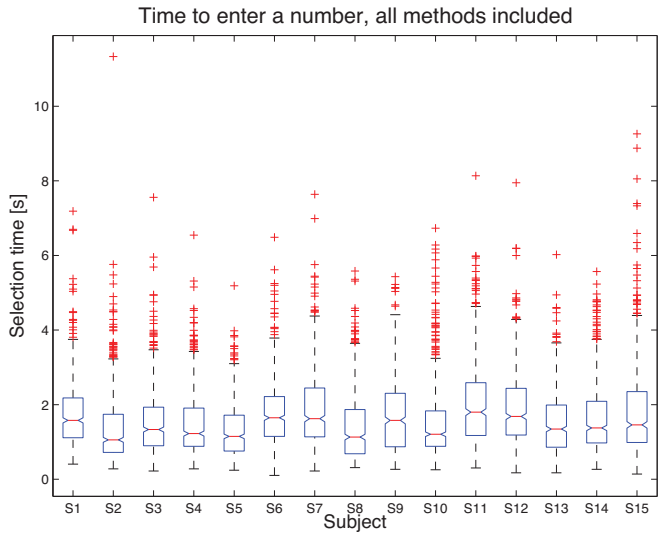

Figure 5.16. Time for selecting one number when all interaction methods are included.

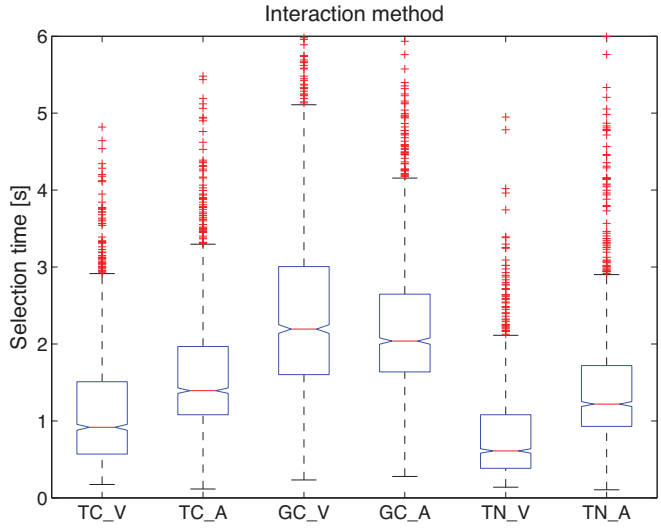

Figure 5.17. Boxplots of selection times for each interaction method.

and 3.08 (GC_A). The mean of the distance from the center across all participants was 74.1 (≈18 cm) (GC_V), 69.3 (≈17 cm) (GC_A), calculated from the pixel values (x, y), received from the Kinect sensor. The difference between audio and visual feedback was significant ($\chi^2 = 32.18$, $p < .01$), but only about 1 cm.

As expected, the numpad visual touch screen (TN_V) was the fastest, because people are very familiar with the layout and accustomed to using it. The touch screen with the circular menu and the visual feedback (TC_V) was second fastest. The touch screen interaction with audio (TN_A and TC_A) performed reasonably well and close to each other. It can be hypothesized that with extended use the performance of the circular menu (TC) will approach that of the numpad menu (TN). Now, 11 of the 15 participants said that the numpad (TN) was more familiar, three considered them equal, and only one said that circular menu felt more familiar.

The gesture interactions (GC) were two times slower than the reference (TN_V), but it is still functional especially for audio-only use(GC_A). Although gesture interaction (GC) is the slowest, its performance should not be directly compared to the touch screen control (TC, and TN). The implementation of the gesture interface can easily be improved to make it faster. Of more interest is how the participants performed with the audio and visual counterparts using the same control method. Differences are seen when comparing the results between the audio and the visual counter parts. When using gesture control, 11 participants were faster with audio (GC_A) than with visual feedback (GC_V). Only one participant performed faster with audio feedback in the touch screen control (TC and TN).

After the experiment, participants were asked if they felt that audio was faster than the visual counterpart. Ten participants said that audio (GC_A) was faster than the visual counter part (GC_V), which is quite close to the results. Eight participants responded that audio (TC_A) was faster than using the touch screen control with the visual circular menu (TC_V). This is surprising, because audio is clearly slower, as seen in Figure 5.17, and the results show that only one participant was actually faster with audio feedback. The visual numpad outperformed the audio counterpart, and only four participants stated that audio (TN_A) was faster than visual (TN_V).

5.6 Application: Funkyplayer

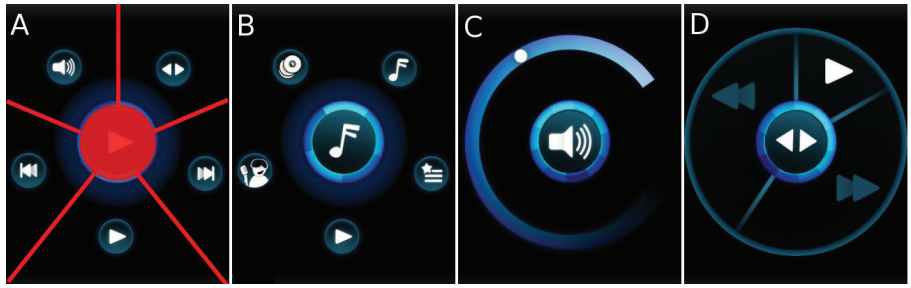

Figure 5.18. a) Sectors defining the menu items in ”now playing” menu, b) the main menu, c) the volume menu, d) the seek menu

The Funkyplayer application was created to integrate and further test visual menus with previously studied auditory menu browsing techniques. Building a stand-alone application was critical for testing the interface in real devices and real use scenarios, e.g., while biking, and necessary for further research. Publication VI continued the previous work introduced in Publications III and IV by introducing improvements to the auditory menu and seamless visual representation. Essentially, Publication VI explored the possibility of combining visual and auditory menus.

Funkyplayer is the first real application that utilizes novel techniques to create an effective eyes-free user interface. Although the interface and menus are designed in terms of audio, it is still pleasing and usable with visual feedback and offers functionality similar to common visual interfaces. Funkyplayer is a program that is used to control the music library of an iPod Touch or iPhone. It performs all the basic functions of a music player, such as browsing and selecting songs, artists, albums, and playlists. Moreover, music playback can be controlled by pausing the music, changing to the next or previous track, adjusting the volume, and winding and rewinding songs. For all controls, Funkyplayer uses egocentric circular menus, which are easy and intuitive to use in the visual and auditory modes.

Funkyplayer was built to demonstrate the possibilities of auditory menus, especially that audio feedback can be efficient in menu browsing and suitable for mobile devices. The music player is only one example of many applications that could benefit from design concepts that can be used without looking at them. This type of interface can be used in navigation and entertainment systems in cars, where it is crucial to keep the eyes on the

road, or in public touch screens where they would improve accessibility for visually impaired users.

Funkyplayer uses gesture and touch screen inputs described in Sections 5.3 and 5.4, but incorporated new features facilitating the eyes-free use of auditory menus. For example, the buttonless gesture mode included a selection gesture and a locking gesture which allows the user ,e.g., to take the iPhone out of the pocket while biking, quickly unlock it, browse and select a new album to play, lock the device again, and return it to the pocket. All of this can be done easily without looking at the device, and even while wearing a pair of thick gloves. The selection gesture is a quick downward motion for selecting a menu item. The motion is performed in the direction of the gravity vector and it is not easily triggered accidentally with any other movement.

5.6.1 Evaluation

Twelve participants completed a user experiment where an iPod Touch running Funkyplayer application was used as the test device. The auditory menu was reproduced with Sennheiser HDR HD-595 headphones connected to the iPod audio output. The screen of the iPod was used to display the visual menu. Input gestures were recognized either with the touch screen or by accelerometers embedded in the device.

The experiment consisted of three tasks using auditory and visual menus measuring time and accuracy. The interaction methods were a touch screen with an auditory menu (TA), touch screen with visual menu (TV), and gesture interaction with auditory menu (GA). The task with each interaction method was the same: finding and selecting ten songs from a list of 147 song names.

After each task, participants filled a System Usability Scale (SUS) [17] questionnaire and answered an open-ended question about negative and positive aspects of the used interaction method. The participants were instructed not to evaluate the features of the music player itself, but the used interaction method and the menu in general. In addition to the SUS questionnaire, the participants filled a short questionnaire for background information and an evaluation of the interaction methods.

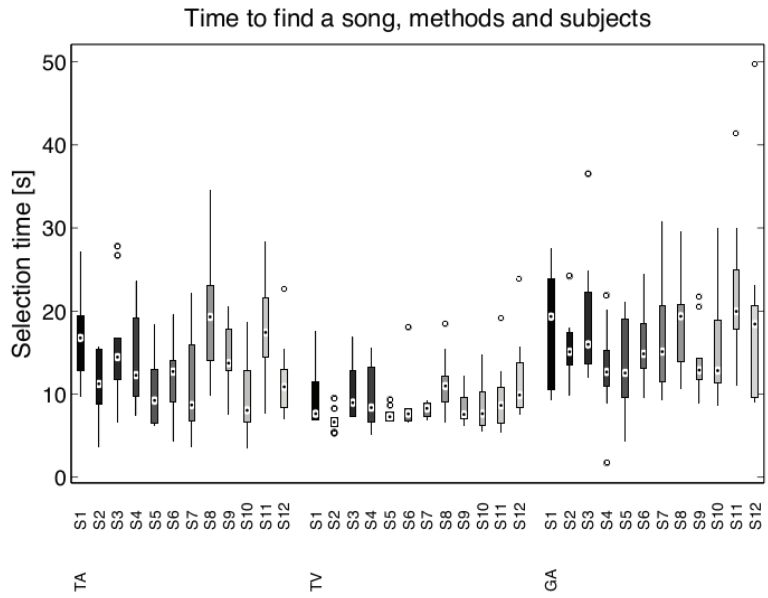

Figure 5.19. Time for selecting one song for each participant and each interaction method.

Results and analysis

Because the distribution of all raw selection times was positively skewed, the median selection times of the names were compared with a non-parametric one-way analysis of variance. In Figure 5.19, participants marked as S1–S6 started with the auditory menu (TA), and the ones marked as S7–S12 started with the visual menu (TV). There was no significant difference in the selection times between the two groups ($\chi^2 = 2.0037$, $p = .1569$), and the individual differences were found to be large.

The median selection times of the three interaction methods are shown in Figure 5.20 and in Table 5.4. The differences between rank means of tasks were also analyzed with the Kruskal-Wallis procedure. The rank means differ significantly ($\chi^2 = 108.32$, $p < .0001$). Post-hoc analysis using Tukey's least significant difference criterion ($p < .05$) of the three conditions shows a difference between all cases (TA, TV, GA). The percentage of correct selections for each interaction method is also shown in Table 5.4.

At the end of each task (TV, TA, GA), participants evaluated the experience with a SUS questionnaire, rating the system features on a 5-point Likert scale. Furthermore, participants gave free form feedback about the negative and the positive aspects of the system after each task. The SUS achieved good internal reliability (Cronbach's α = 0.77 (TV), 0.79 (TA),

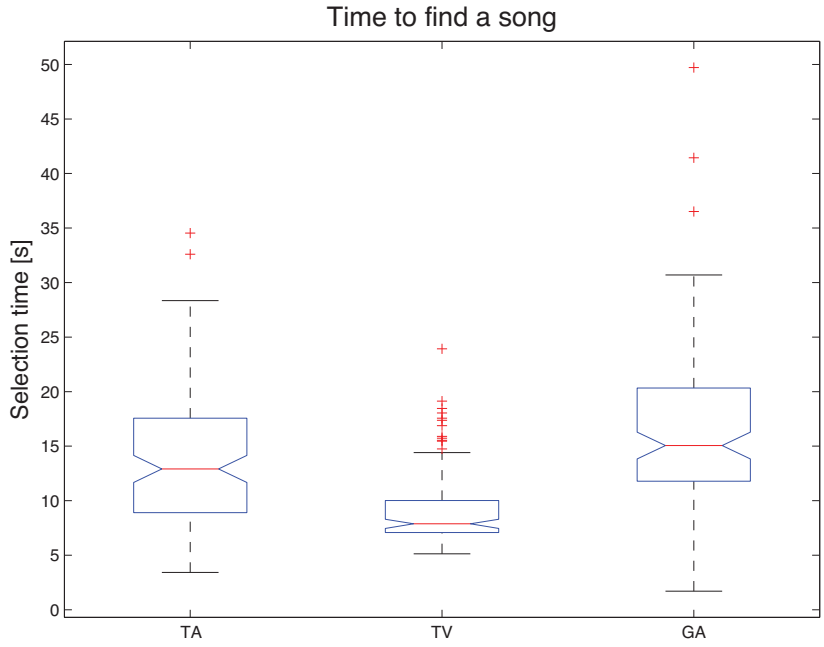

Figure 5.20. Time for selecting one song with three interaction methods (TA, TV, GA).

0.84 (GA)).

The SUS scores are summarized in Table 5.4. A 2 times 3 (order times interaction method) ANOVA was conducted on the mean SUS scores, and no significant effect of the starting order was found ($F(1, 35) = 1.73$, $p = .2178$). Therefore, all SUS scores were combined without considering the presentation order and the one-way ANOVA found a significant main effect of the interaction method on SUS scores $F(2, 35) = 16.42$, $p < .0001$. Post-hoc Tukey multiple comparison of means revealed that the SUS scores of the GA differed significantly from TV ($p < .001$) and TA

Table 5.4. Results of the user experiment.

Time and accuracy	TA	TV	GA
Correct selections [%]	97.5	90.8	94.2
Median selection times [s]	12.91	7.87	15.06
SUS scores	TA	TV	GA
Auditory menu first	76.25	73.34	50.21
Visual menu first	75.00	60.83	41.25
Total mean SUS score	**75.63**	**67.87**	**45.83**

($p < .001$), but TV and TA did not have a significant difference ($p = .2747$). To conclude, the overall SUS scores positioned the system usability between *Good* and *Excellent* (TV), *OK* and *Good* (TA), and *Poor* and *OK* (GA) [5].

The selection times for the three interaction methods varied significantly. The touch screen with the auditory menu (TA) was on average 5 s slower than the visual menu (TV). However, some participants were almost as fast with TA as with TV, as seen in Figure 5.19. Both touch screen interaction methods (TA, TV) can be considered relatively fast and accurate to use. The SUS scores suggest that the touch screen interfaces are usable and can even be fun to use. People are not accustomed to auditory and eyes-free interfaces, which may cause confusion for some users. The SUS scores for TA and TV were quite close, especially in the group that started the experiment with the auditory menu (see Table 5.4). The gesture interaction (GA) received the poorest results and SUS scores. In addition, it was 7 s slower than the visual menu (TV), which indicates that there is room for improvement. However, some participants definitely liked it, and the performance of some participants was closer to touch screen (TA) performance. The speed and accuracy is still good, when taking into account that no visual feedback was given. All participants were also asked to evaluate on a scale from 1 (very hard) to 5 (very easy) how easy switching from visual to audio or vice versa was. The average score was 3.84, which suggests that switching between the modalities is easy.

The participants gave written positive and negative feedback after using each interaction method. Especially both touch screen methods (TV, TA) received praises, and the gesture interaction (GA) got suggestions for further improvement.

The touch screen with the visual menu (TV) received many positive comments, such as the interface being *"easy" or "easy to learn"* (7 times) and "navigation was generally intuitive," "the alphabetical circular menu is logical and natural." The most common negative comments were: *"text is behind my thumb"* (6 times), *"letters (in the alphabet) are quite close to each other"* (3 times), and *"need to be precise when selecting"* (4 times).

The touch screen with the auditory menu (TA) was found to be *"easy to use"* (7 times) and also *"very fast to use," "fast, fun and very precise."* Four participants highlighted the eyes-free use: *"surprisingly easy to use without looking, after you learn the application logic," "can be used with eyes closed."* Four participants found the selection by releasing the finger

cumbersome: *"making a selection by lifting the finger is maybe not the most convenient way."* Furthermore, the participants requested a "where am I" functionality for use when lost in the menu structure, although it was already implemented.

Gesture interaction with the auditory menu (GA) received positive feedback that shows that some participants felt confident with the interaction method: *"circular browsing is fluent," "fast browsing is easy"*. The gesture recognition was also complimented: *"the gesture detection is accurate"* (3 times). Negative comments pointed out that the tilt angle for the center area should be larger: *"finding the center was hard"* (5 times). The selection gesture should be improved: *"selection is not made every time"* (6 times) and the ergonomics of the device and the gesture should be revised: *"the hand can get tired when using longer"* (6 times).

5.7 Discussion and lessons learned

All the studies suggest that gesture-based interaction with auditory menus can provide a fast and eyes-free way to control devices, which is especially beneficial when visual attention should be focused elsewhere. The results of the studies suggest ways to improve the eyes-free gesture interfaces and gives general design recommendations, which are discussed below.

The use of ergonomic gestures and a device that fits in the hand is important. The studies and user comments pointed out that the slim form factor of the current smart phones is not optimal for these kinds of circular gestures, and effortless use also requires more time to learn. A device shaped like a tube or a joystick-style device held with closed fingers would be more ergonomic to use with the circular wrist gesture, as illustrated in Figure 5.6. Designing and building a designated device for possible future studies would be interesting. In addition to a better form factor, a dedicated device would allow better placement of buttons or attaching pressure sensors for detecting squeezing.

Different selection methods and gestures were already used in different studies ranging from using a button, or releasing a finger from a touch screen to downward hand motion. Although studies show that the downward selection gesture is usable with hand-held devices and free-hand interaction, still the rehearsing time can vary between individuals, and it is not as fast to learn as pushing a button, squeezing, or using a "binary

gesture". A "binary gesture", (like pinching two fingers together) can be more accurate in free-hand interaction and proprioception (the sense of the relative position of neighboring fingers in this context) can give natural feedback for more a successful gesture. For hand-held devices, buttons or squeezing are a faster and more reliable selection method.

The hand getting tired is a problem with all gesture interfaces where the arm needs to be held up for longer periods. Gesture interfaces should be designed so that the gestures are small and the hand is kept as low as possible to ease the effort made by the muscles. Furthermore, a horizontal hand movement is especially suitable for circular motion because it is less tiring than a vertical one.

The eyes-free touchscreen interaction has proved effective in studies described earlier in this chapter. The circular gesture around the center is fast to learn and robust to use, but because many touch screens do not have physical borders, the finger can accidentally slide off the screen. Sliding off the surface should be detected and the user be made aware of it. The selection method for touch screen interaction can be just releasing the finger when it is above a menu item. However, this can cause accidental finger lifts and thus it is better suited for more experienced users. When using a tap (or a double tap) for selection, the finger can always be released, and tapping anywhere in the screen selects the active item. A tap (or a double tap) is a slightly slower selection method, but conveniently allows releasing the finger anytime without making a selection. Additionally, a second-finger tap can be used for item selection [55].

The circular auditory menu in general has proved to be relatively fast and accurate with all the tested interaction methods. The design of the auditory menu has been improved during the course of the studies presented in this thesis. Giving feedback to the user helps especially when using an unfamiliar menu structure. Implementing a "Where am I" functionality that can be always used to verify the position in the menu structure is beneficial. Care should be taken to play only one menu item name at a time, because simultaneous menu item names can confuse a novice user. Further work should investigate how better feedback can be implemented by attaching continuous audible information to the menu levels or items. This can provide information about location in the menu hierarchy [105]. Also the traditional text-to-speech approach can be changed in different ways, e.g., by using whispered sounds for unavailable menu items, which in the visual domain would be grayed out [52]. Using continuous

sounds for certain menu items or repeating (or trailing off an echo) the name of the current menu item can be helpful.

The introduced dynamic spreading method extends the number of items that can be efficiently browsed in one single menu level and it also enables fast transition to far away position in the menu. The long lists used in the two experiments described earlier in this chapter contained about 150 names, which is more than previous eyes-free interfaces have been capable of [55]. However in real application, contact list or music library can hold thousands of names and multi-layered menus can still be needed. Although dynamic menu item spreading might scale up to thousand names, testing its upper limits and further improvements remain as future work.

The results of the tests can be used to improve applications and devices that would benefit from the eyes-free and gesture interaction. One application possibility is a glove or a ring with embedded orientation sensors that can be used discretely and remotely to control basic functionalities of a mobile phone (see Figure 5.21). Interacting with a smartphone without taking it out of the pocket can be useful in cold or dirty environments, e.g, while snowboarding. It is also possible to construct a small multifunctional device consisting only of internal rotation-sensing devices, e.g., accelerometers. Such a robust device without a visual display could perform all the controls of a simple mobile phone.

Front and back facing cameras may also be used to interact with a mobile phone without picking it up or touching the phone. Quick interaction is possible when the phone is on a table or in a car dock. Figure 5.22 illustrates two concepts for IVS. A camera mounted on a ceiling or dashboard of a car may be usedto detect gestures. Touch screen applications could be designed to be used in an eyes-free mode, which allows the driver to concentrate on the traffic. Furthermore, a gesturally augmented desktop environment could work as assistive technology for visually impaired users. A hand can be placed on a predefined virtual device (e.g., radio) which triggers an audible notification. The device could be activated with a selection gesture and controlled with circular auditory menu. This kind of interface could be used to control radio, phone, or computer applications.

Circular interaction has potential with other input methods. Auditory menus and visual circular menus could be accessed with eye movements. This might be a good solution for controlling head-mounted displays and especially for people with visual or other impairments. Also, the suitabil-

ity of circular auditory menus for efficient access to text-based information and typing should be researched.

Figure 5.21. Gesture and touch surface control with an auditory menu can be used in devices without a visual display or to remotely control a mobile phone. A touch surface can be attached to a sleeve. Accelerometers attached to a glove, next to the back of the hand, can detect small wrist movements while the arm is relaxed and pointing down.

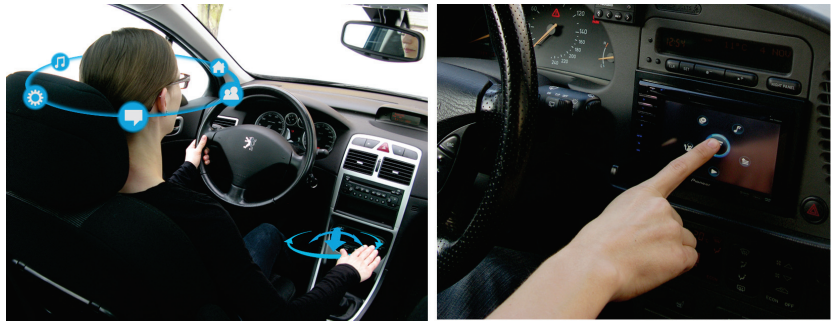

Figure 5.22. Car infotainment systems will benefit from menus that are designed to be used without looking at them. Free-hand gestures recognized with a ceiling camera could be used to control infotainment systems (left) or touch gestures can be used on a touch screen that can be turned off while moving (right). An auditory menu can be reproduced with loudspeakers of a car.

6. Summary

6.1 Main results of the thesis

A summary of the main findings and results in this thesis are listed as follows:

- The perceived shape and size of a space can be changed by applying different reverberation times in different directions using multiple reverberation systems. The prototyped system was utilized in live performances and to enhance lecture room acoustics.

- An experimental opera production explored how gesture control can be utilized in live performances for natural timing of events and enriching artistic expression. The performer-based interaction is useful when user controlled media is directly mapped to gestures and when detailed nuances of movement are hard for a technician controlling the media to follow. Performing arts with more room for improvisation can make the most of the media controlled by performers.

- Three gestural control methods for controlling circular auditory menus are proposed. All control methods are accurate and fast enough for efficient eyes-free use. The circular control metaphor is used with all control methods and it enables the access to the same auditory menu.

- Several eyes-free browsing methods for auditory menus are presented. In particular the proposed dynamic menu item spreading is efficient and accurate for large auditory menus. Combining eyes-free interfaces and a browsing method with a dynamically adjustable target size of the menu items allows the use of large menus with intuitive and easy access.

- A interface paradigm is presented for interoperable auditory and visual menus. The same control logic for both visual and auditory domains may facilitate switching to eyes-free use when needed and may improve accessibility for visually impaired users.

- Auditory menus can be as fast or slightly faster than visual menus even with the participant's full attention on the task. The proposed free-hand gesture method's performance and accuracy with an auditory interface is the same or even slightly better than the visual one.

6.2 Future work

The research presented in this thesis can be continued in following ways:

- Further development of advanced auditory menus enabling efficient eyes-free access, especially layouts and properties suitable for efficient text entry and accessing text-based or audio-based information should be done.

- Investigating natural gesture control and input methods for auditory menus. Input methods where small gestures can be performed discretely or concealed from others should be investigated.

- Study on how identical control gestures in visual and auditory menus affect learning of the menu layout and facilitate the switch between the modalities should be continued.

- Further studies should be done on how ubiquitous sensors embedded, e.g., in clothing in the performing arts can be utilized in natural manner by the performer with a small cognitive load.

- Supplementary investigation on how electronically modified acoustics can be utilized in performances and how practice rooms can be enhanced with an electroacoustic system should be carried out. Stress should be laid on how the reverberation algorithm should be implemented and the loudspeaker and the microphone setup built for realistic results.

Bibliography

[1] M. Akamatsu, I. MacKenzie, and T. Hasbrouq. A comparison of tactile, auditory, and visual feedback in a pointing task using a mouse-type device. *Ergonomics*, 38:816–827, 1995.

[2] D. Ashbrook, P. Baudisch, and S. White. Nenya: subtle and eyes-free mobile input with a magnetically-tracked finger ring. In *Proceedings of CHI '11*, pages 2043–2046, New York, NY, USA, 2011. ACM.

[3] G. Bailly, D.-B. Vo, E. Lecolinet, and Y. Guiard. Gesture-aware remote controls: guidelines and interaction technique. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 263–270, New York, NY, USA, 2011. ACM.

[4] M. Baldauf, S. Zambanini, P. Fröhlich, and P. Reichl. Markerless visual fingertip detection for natural mobile device interaction. In *Proceedings of MobileHCI '11*, pages 539–544, Stockholm, Sweden, 2011. ACM.

[5] A. Bangor, P. Kortum, and J. Miller. Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of Usability Studies*, 4(3):114–123, 2009.

[6] D. R. Begault. Perceptual effects of synthetic reverberation on three-dimensional audio systems. *Journal of the audio engineering society*, 40(11):895–904, 1992.

[7] D. R. Begault. Virtual acoustics, aeronautics, and communications. *Journal of the audio engineering society*, 46(6):520–530, 1998.

[8] D. R. Begault. *3-D sound for virtual reality and multimedia*. National aeronautics and space administration, 2000.

[9] M. Blattner, D. Sumikawa, and R. Greenberg. Earcons and icons: Their structure and common design principles. *Journal Human Computer Interaction*, 4(1):11–44, 1989.

[10] J. Blauert. *Spatial hearing: The psychophysics of human sound localization*. Cambridge, MA: MIT Press, 1983.

[11] R. A. Bolt. "Put-that-there": Voice and gesture at the graphics interface. In *Proceedings of SIGGRAPH '80*, pages 262–270, New York, NY, USA, 1980. ACM.

[12] M. N. Bonner, J. T. Brudvik, G. D. Abowd, and W. K. Edwards. No-look notes: accessible eyes-free multi-touch text entry. In *Proceedings of Pervasive '10*, pages 409–426, 2010.

[13] M. Boone, E. Verheijen, and P. van Tol. Spatial sound-field reproduction by wave-field synthesis. *Journal of the audio engineering society*, 43(12):1003–1012, 1995.

[14] S. Brewster, F. Chohan, and L. Brown. Tactile feedback for mobile interactions. In *Proceedings of the SIGCHI conference on human factors in computing systems*, CHI '07, pages 159–162, New York, NY, USA, 2007. ACM.

[15] S. A. Brewster, J. Lumsden, M. Bell, M. Hall, and S. Tasker. Multimodal "eyes-free" interaction techniques for wearable devices. In *Proceedings of CHI '03*, pages 463–480. ACM, 2003.

[16] S. A. Brewster, P. Wright, and A. Edwards. A detailed investigation into the effectiveness of earcons. In *Proceedings of ICAD '92*, pages 471–498, 1992.

[17] J. Brooke. SUS: a quick and dirty usability scale. *Usability Evaluation in Industry*, pages 189–194, 1996.

[18] C. Cadoz. *Les realites virtuelles*. Dominos, Flammarion, 1994.

[19] C. Cadoz and M. Wanderley. Gesture-music. *Wanderley, M., Battier, M. (eds.) Trends in gestural control of music (Edition électronique)*, pages 71–93, 2000.

[20] J. Callahan, D. Hopkins, M. Weiser, and B. Shneiderman. An empirical comparison of pie vs. linear menus. In *Proceedings of CHI '88*, pages 95–100, New York, NY, USA, 1988. ACM.

[21] J. Cassell. *A framework for gesture generation and interpretation*, pages 191–215. Cipolla, R. and Pentland, A. (eds.), 1998.

[22] C. I. Cheng and G. H. Wakefield. Introduction to head-related transfer functions (HRTFs): Representations of HRTFs in time, frequency, and space. *Journal of the audio engineering society*, 49(4):231–249, 2001.

[23] M. Cohen and L. F. Ludwig. Multidimensional audio window management. In S. Greenberg, editor, *Computer-supported cooperative work and groupware*, pages 193–210. Academic Press Ltd., London, UK, UK, 1991.

[24] E. Costanza, S. A. Inverso, R. Allen, and P. Maes. Intimate interfaces in action: assessing the usability and subtlety of emg-based motionless gestures. In *Proceedings of CHI '07*, pages 819–828, New York, NY, USA, 2007. ACM.

[25] K. Crispien, K. Fellbaum, A. Savidis, and C. Stephanidis. 3D-auditory environment for hierarchical navigation in non-visual interaction. In *Proceedings of ICAD '96*, pages 18–21, 1996.

[26] A. Crossan, J. Williamson, S. A. Brewster, and R. Murray-Smith. Wrist rotation for interaction in mobile contexts. In *Proceedings of MobileHCI'08*, pages 435–438. ACM, 2008.

[27] C. Dicke, S. Deo, M. Billinghurst, N. Adams, and J. Lehikoinen. Experiments in mobile spatial audio-conferencing: key-based and gesture-based interaction. In *Proceedings of MobileHCI '08*, pages 91–100, New York, NY, USA, 2008. ACM.

[28] C. Dicke, K. Wolf, and Y. Tal. Foogue: eyes-free interaction for smartphones. In *Proceedings of MobileHCI '10*, pages 255–248, 2010.

[29] R. O. Duda and W. M. Martens. Range dependence of the response of a spherical head model. *Journal of the acoustical society of america*, 5(104):3048–3058, 1998.

[30] A. Farina. Simultaneous measurement of impulse response and distortion with a swept-sine technique. In *Proceedings of 108th convention of the AES*, volume 48, page 350, April 2000.

[31] C. Frauenberger, V. Putz, R. Höldrich, and T. Stockman. Interaction patterns for auditory user interfaces. In *Proceedings of ICAD '05*, pages 154–160, 2005.

[32] C. Frauenberger, T. Stockman, V. Putz, and R. Holdrich. Mode independent interaction pattern design. In *Information visualisation*, pages 24–30, 2005.

[33] B. Frey, C. Southern, and M. Romero. Brailletouch: mobile texting for the visually impaired. In *Proceedings of UAHCI '11*, pages 19–25, Berlin, Heidelberg, 2011. Springer-Verlag.

[34] N. Friedlander, K. Schlueter, and M. Mantei. Bullseye! when Fitts' law doesn't fit. In *Proceedings of CHI '98*, pages 257–264, New York, NY, USA, 1998. ACM Press/Addison-Wesley Publishing Co.

[35] N. Friedlander, K. Schlueter, and M. M. Mantei. A bullseye menu with sound feedback. In *Proceedings of HCI '97*, pages 379–382, 1997.

[36] G. Furnas. Generalized fisheye views. In *Proceedings CHI '86*, pages 16–23. ACM, 1986.

[37] W. Gaver. Auditory icons: using sound in computer interfaces. *Journal Human Computer Interaction*, 2(2):167–177, 1986.

[38] M. Gerzon. Periphony: With-height sound reproduction. *Journal of the audio engineering society*, 21(1):2–10, 1973.

[39] T. Guerreiro, P. Lagoá, H. Nicolau, D. Gonçalves, and J. Jorge. From tapping to touching: making touch screens accessible to blind users. *IEEE Multimedia*, 15(4):48–50, 2008.

[40] S. Gustafson, D. Bierwirth, and P. Baudisch. Imaginary interfaces: spatial interaction with empty hands and without visual feedback. In *Proceedings of UIST '10*, pages 3–12, New York, NY, USA, 2010. ACM.

[41] H. Haas. The influence of a single echo on the audibility of speech. *Journal of the audio engineering society*, 20(2):146–159, 1972.

[42] A. Härmä, J. Jakka, M. Tikander, M. Karjalainen, T. Lokki, J. Hiipakka, and G. Lorho. Augmented reality audio for mobile and wearable appliances. *Journal of the audio engineering society*, 52(6):618–639, 2004.

[43] B. L. Harrison, K. P. Fishkin, A. Gujar, C. Mochon, and R. Want. Squeeze me, hold me, tilt me! an exploration of manipulative user interfaces. In *Proceedings of CHI '98*, pages 17–24, New York, NY, USA, 1998. ACM Press/Addison-Wesley Publishing Co.

[44] T. Hermann, T. Henning, and H. Ritter. Gesture desk – an integrated multi-modal gestural workplace for sonification. *Lecture notes in computer science*, 2915:103–104, 2004.

[45] E. Hoggan, A. Crossan, S. A. Brewster, and T. Kaaresoja. Audio or tactile feedback: which modality when? In *Proceedings of CHI '09*, pages 2253–2256, New York, NY, USA, 2009. ACM.

[46] C. Holz, T. Grossman, G. Fitzmaurice, and A. Agur. Implanted user interfaces. In *Proceedings of CHI '12*, pages 503–512, New York, NY, USA, 2012. ACM.

[47] iPod shuffle: portable music player. http://www.apple.com/, Last checked 1.9.2012.

[48] D. Jagdish and M. Gupta. Sonic grid: an auditory interface for the visually impaired to navigate gui-based environments. In *Proceedings of the IUI '08*, pages 337–340, 2008.

[49] M. Jain and R. Balakrishnan. User learning and performance with bezel menus. In *Proceedings of CHI '12*, pages 2221–2230, New York, NY, USA, 2012. ACM.

[50] C. Jayant, C. Acuario, W. Johnson, J. Hollier, and R. Ladner. V-braille: Haptic braille perception using a touch-screen and vibration on mobile phones. In *Proceedings of ACCETS 2010*, pages 295–296, 2010.

[51] M. Jeon, B. K. Davison, M. A. Nees, J. Wilson, and B. N. Walker. Enhanced auditory menu cues improve dual task performance and are preferred with in-vehicle technologies. In *Proceedings of AutomotiveUI '09*, pages 91–98, 2009.

[52] M. Jeon, S. Gupta, B. Davison, and B. Walker. Auditory menus are not just spoken visual menus: A case study of "unavailable" menu items. In *CHI EA '10*, pages 3319–3324, 2010.

[53] M. Jeon and B. N. Walker. "Spindex": Accelerated initial speech sounds improve navigation performance in auditory menus. In *Proceedings of HFES 2009*, 2009.

[54] D. Kammer, J. Wojdziak, M. Keck, R. Groh, and S. Taranko. Towards a formalization of multi-touch gestures. In *Proceedings of ITS '10*, pages 49–58, New York, NY, USA, 2010. ACM.

[55] S. K. Kane, J. P. Bigham, and J. O. Wobbrock. Slide rule: making mobile touch screens accessible to blind people using multi-touch interaction techniques. In *Proceedings of ASSETS '08*, pages 73–80, 2008.

[56] S. K. Kane, J. O. Wobbrock, and R. E. Ladner. Usable gestures for blind people: understanding preference and performance. In *Proceedings of CHI '11*, pages 413–422, New York, NY, USA, 2011. ACM.

[57] M. Karam and M. C. Schraefel. A taxonomy of gestures in human computer interactions. Technical report, Electronics and computer science, University of Southampton, 2005.

[58] M. Karjalainen, T. Mäkipatola, A. Kanerva, and A. Huovilainen. Virtual air guitar. *Journal of the audio engineering society*, 54(10):964–980, 2006.

[59] S. Keates and P. Robinson. The use of gestures in multimodal input. In *Proceedings of ASSETS '98*, pages 35–42, New York, NY, USA, 1998. ACM.

[60] J. Kela, P. Korpipää, M. Jani, S. Kallio, G. Savino, L. Jozzo, and D. Marca. Accelerometer-based gesture control for a design environment. *Personal Ubiquitous Computing*, 10(5):285–299, July 2006.

[61] J. Kim, J. He, K. Lyons, and T. Starner. The gesture watch: a wireless contact-free gesture based wrist interface. In *Proceedings of 11th IEEE international symposium on wearable computers*, pages 15–22, 2007.

[62] Kinect. http://www.xbox.com/en-us/kinect, Last checked 1.9.2012.

[63] M. Kobayashi and C. Schmandt. Dynamic soundscape: mapping time to space for audio browsing. In *Proceedings of human factors in computing systems*, pages 194–201, 1997.

[64] G. Kramer, B. Walker, T. Bonebright, P. Cook, J. Flowers, N. Miner, and J. Neuhoff. Sonification report: Status of the field and research agenda. Technical report, International community for auditory display, 1999.

[65] S. Kuhn. Extended presence: The instrumental(ised) body in andré werner's marlowe: The jew of malta. *International journal of performance arts and digital media*, 2(3):221–236, 2006.

[66] F. C. Y. Li, D. Dearman, and K. N. Truong. Virtual shelves: interactions with orientation aware devices. In *Proceedings of the 22nd annual ACM symposium on user interface software and technology*, pages 125–128, Victoria, BC, Canada, 2009. ACM.

[67] A. Löcken, T. Hesselmann, M. Pielot, N. Henze, and S. Boll. User-centred process for the definition of free-hand gestures applied to controlling music playback. *Multimedia systems*, 18:15–31, 2012.

[68] T. Lokki, J. Nummela, and T. Lahti. An electro-acoustic enhancement system for rehearsal rooms. In *Proceedings of EAA Symposium on architectural acoustics*, 2000.

[69] T. Lokki, J. Pätynen, T. Peltonen, and O. Salmensaari. A rehearsal hall with virtual acoustics for symphony orchestras. In *Proceedings of the AES 126th international convention*, page paper nr. 7695, 2009.

[70] G. Lorho, J. Marila, and J. Hiipakka. Feasibility of multiple non-speech sounds presentation using headphones. In *Proceedings of ICAD '01*, pages 32–37, 2001.

[71] S. MacKenzie and S. Castellucci. Reducing visual demand for gestural text input on touchscreen devices. In *Proceedings of CHI EA '12*, pages 2585–2590, New York, NY, USA, 2012. ACM.

[72] D. Malham and A. Myatt. 3-D sound spatialization using ambisonic techniques. *Computer Music Journal*, 19(4):58–70, 1995.

[73] G. Marentakis and S. A. Brewster. Effects of feedback, mobility and index of difficulty on deictic spatial audio target acquisition in the horizontal plane. In *Proceedings of CHI '06*, pages 359–368, 2006.

[74] Z. Mo, J. P. Lewis, and U. Neumann. Smartcanvas: a gesture-driven intelligent drawing desk system. In *Proceedings of IUI '05*, pages 239–243, New York, NY, USA, 2005. ACM.

[75] Mrmr. http://poly.share.dj/projects/, Last checked 1.9.2012.

[76] F. Mueller and M. Karau. Transparent hearing. In *Proceedings of CHI EA '02*, pages 730–731, New York, NY, USA, 2002. ACM.

[77] A. Neustein. *Advances in speech recognition: mobile environments, call centers and clinics*. Springer, 1st edition, 2010.

[78] M. Nielsen, M. Störring, T. B. Moeslund, and E. Granum. A procedure for developing intuitive and ergonomic gesture interfaces for man-machine interaction. Technical report, Aalborg University, 2003.

[79] I. Oakley and J. Park. A motion-based marking menu system. In *Proceedings of CHI '07*, pages 2597–2602. ACM, 2007.

[80] I. Oakley and J.-S. Park. Designing eyes-free interaction. In *Proceedings of HAID'07*, pages 121–132, Berlin, Heidelberg, 2007. Springer-Verlag.

[81] M. Orth. Interface to architecture: integrating technology into the environment in the brain opera. In *Proceedings of the 2nd conference on Designing interactive systems: processes, practices, methods, and techniques*, pages 265–275, 1997.

[82] A. Palmer, T. Shackleton, and D. McAlpine. Neural mechanisms of binaural hearing. *Acoustical science and technology*, 23(2):61–68, 2002.

[83] J. Paradiso. The brain opera technology: New instruments and gestural sensors for musical interaction and performance. *Journal of new music research*, 28(2):130–149, 1999.

[84] C. Park, P. Chou, and Y. Sun. A wearable wireless sensor platform for interactive dance performances. In *Proceedings of PerCom '06*, pages 52–59, 2006.

[85] K. Partridge, S. Chatterjee, V. Sazawal, G. Borriello, and R. Want. Tilttype: accelerometer-supported text entry for very small devices. In *Proceedings of UIST '02*, pages 201–204, New York, NY, USA, 2002. ACM.

[86] J. Pasquero, S. J. Stobbe, and N. Stonehouse. A haptic wristwatch for eyes-free interactions. In *Proceedings of CHI '11*, pages 3257–3266, New York, NY, USA, 2011. ACM.

[87] PD. Pure data. http://puredata.info/, Last checked 1.9.2012.

[88] C. A. Pickering, K. J. Burnham, and M. J. Richardson. A research study of hand gesture recognition technologies and applications for human vehicle interaction. In *Proceedings of automotive electronics*, pages 1–15, 2007.

[89] M. Pielot, B. Poppinga, W. Heuten, and S. Boll. A tactile compass for eyes-free pedestrian navigation. In P. Campos, N. Graham, J. Jorge, N. Nunes, P. Palanque, and M. Winckler, editors, *Human-computer interaction – INTERACT 2011*, volume 6947 of *Lecture Notes in Computer Science*, pages 640–656. Springer Berlin / Heidelberg, 2011.

[90] C. S. Pinhanez and A. F. Bobick. It/I: an experiment towards interactive theatrical performances. In *Proceedings of CHI '98*, pages 333–334, 1998.

[91] A. Pirhonen, S. A. Brewster, and C. Holguin. Gestural and audio metaphors as a means of control for mobile devices. In *Proceedings of CHI '02*, pages 291–298. ACM, 2002.

[92] M. Puckette. Pure data: another integrated computer music environment. In *Second intercollege computer music concerts*, pages 37–41, 1996.

[93] V. Pulkki. *Spatial sound generation and perception by amplitude panning tecniques*. PhD thesis, Helsinki university of technology, 2001.

[94] V. Pulkki, M.-V. Laitinen, and V. Sivonen. HRTF measurements with a continuously moving loudspeaker and swept sines. In *Proceedings of 128th Convention of the AES*, volume 58, page 178, March 2010.

[95] M. Rahman, S. Gustafson, P. Irani, and S. Subramanian. Tilt techniques: investigating the dexterity of wrist-based input. In *Proceedings of CHI '09*, pages 1943–1952, New York, NY, USA, 2009. ACM.

[96] V. C. Raykar, R. Duraiswami, and B. Yegnanarayana. Extracting the frequencies of the pinna spectral notches in measured head related impulse responses. *Journal of the acoustical society of america*, 118(1), 2005.

[97] J. Rekimoto. Tilting operations for small screen interfaces. In *Proceedings of UIST*, pages 167–168. ACM, 1996.

[98] B. Rime and L. Schiaratura. Gesture and speech. *Fundamentals of non-verbal behavior*, pages 239–281, 1991.

[99] D. Rocchesso. Audio effects to enhance spatial information displays. In *First international symposium on 3D data processing visualization and transmission*, 2002.

[100] M. Romero, B. Frey, C. Southern, and G. Abowd. Brailletouch: Designing a mobile eyes-free soft keyboard. In *Proceedings of MobileHCI '11, Design competition*, 2011.

[101] A. Savidis, C. Stephanidis, A. Korte, K. Crispien, and K. Fellbaum. A generic direct-manipulation 3D-auditory environment for hierarchical navigation in non-visual interaction. In *Proceedings of ASSETS '96*, pages 117–123. ACM, 1996.

[102] L. Savioja, J. Huopaniemi, T. Lokki, and R. Väänänen. Creating interactive virtual acoustic environment. *Journal of the audio engineering society*, 47(9):675–705, 1999.

[103] N. Sawhney and C. Schmandt. Nomadic radio: Speech and audio interaction for contextual messaging in nomadic environments. *ACM Transactions on computer-human interaction*, 7(3):353–383, 2000.

[104] V. Sazawal, R. Want, and G. Borriello. The unigesture approach. *Mobile-HCI '02*, pages 256–270, 2002.

[105] E. Sikström and J. Berg. Designing auditory display menu interfaces - cues for users current location in extensive menus. In *Proceedings of 126th AES convention*, 5 2009.

[106] Siri. http://www.apple.com/iphone/features/siri.html, Last checked 1.9.2012.

[107] J. Sodnik, C. Dicke, S. Tomažič, and M. Billinghurst. A user study of auditory versus visual interfaces for use while driving. *International journal of human-computer studies*, 66(5):318–332, 2008.

[108] S. Spagnol, M. Geronazzo, and F. Avanzini. Fitting pinna-related transfer functions to anthropometry for binaural sound rendering. In *Proceedings of IEEE international workshop on multimedia signal processing*, pages 194–199, 2010.

[109] F. Sparacino, C. Wren, G. Davenport, and A. Pentland. Augmented performance in dance and theater. *International dance and technology*, 1999.

[110] D. Spelmezan, M. Jacobs, A. Hilgers, and J. Borchers. Tactile motion instructions for physical activities. In *Proceedings of CHI '09*, pages 2243–2252, New York, NY, USA, 2009. ACM.

[111] R. Spence and M. Apperley. Data base navigation: an office environment for the professional. *Behaviour & information technology*, 1(1):43–54, 1982.

[112] R. M. Stanley and B. N. Walker. Lateralizaton of sounds using bone-conduction headsets. In *Proceedings of the annual meeting of the human factors and ergonomics society*, pages 1571–1575, 2006.

[113] R. Stewart, M. Levy, and M. Sandler. 3D interactive environment for music collection navigation. In *Proceedings of DAFX*, pages 13–17, 2008.

[114] L. Tarabella. Handel, a free-hands gesture recognition system. In *Proceedings of computer music modeling and retrieval*, volume 3310 of *Lecture Notes in Computer Science*, pages 139–148. Springer Berlin / Heidelberg, 2005.

[115] F. Tian, L. Xu, H. Wang, X. Zhang, Y. Liu, V. Setlur, and G. Dai. Tilt menu: Using the 3D orientation information of pen devices to extend the selection capability of pen-based user interfaces. In *Proceedings of CHI '08*, pages 1371–1380. ACM, 2008.

[116] H. Tinwala and I. S. MacKenzie. Eyes-free text entry with error correction on touch screen mobile devices. In *Proceedings of NordiCHI '10*, pages 511–520, 2010.

[117] K. Usui, M. Takano, Y. Fukushima, and I. Yairi. The evaluation of visually impaired people's ability of defining the object location on touch-screen. In *Proceedings of ASSETS '10*, pages 287–288, 2010.

[118] Y. Vazquez-Alvarez and S. A. Brewster. Eyes-free multitasking: the effect of cognitive load on mobile spatial audio interfaces. In *Proceedings of CHI '11*, pages 2173–2176, New York, NY, USA, 2011. ACM.

[119] Y. Vazquez-Alvarez, I. Oakley, and S. Brewster. Auditory display design for exploration in mobile audio-augmented reality. *Personal and Ubiquitous Computing Journal*, pages 1–13, 2011.

[120] Y. Vazquez-Alvarez, I. Oakley, and S. Brewster. Eyes-free multitasking: The effect of cognitive load on mobile spatial audio interfaces. In *Proceedings of CHI '11*, pages 2173–2176, 2011.

[121] Voice actions for android. http://www.google.com/mobile/voice-actions/, Last checked 1.9.2012.

[122] J. Wachs, H. Stern, Y. Edan, M. Gillam, C. Feied, M. Smith, and J. Handler. A hand-gesture sterile tool for sterile browsing of radiology images. *Journal of the american medical informatics association*, 15(3):321–323, 2008.

[123] J. P. Wachs, M. Kölsch, H. Stern, and Y. Edan. Vision-based hand-gesture applications. *Communications of the ACM*, 54:60–71, February 2011.

[124] A. Walker and S. A. Brewster. Extending the auditory display space in handheld computing devices. In *Proceedings of second workshop on human computer interaction with mobile devices*, 1999.

[125] A. Walker, S. A. Brewster, D. McGookin, and A. Ng. A diary in the sky: A spatial audio display for a mobile calendar. In *Proceedings of BCS IHM-HCI*, pages 531–540. Springer, 2001.

[126] B. N. Walker, J. Lindsay, A. Nance, Y. Nakano, D. K. Palladino, T. Dingler, and M. Jeon. Spearcons (speech-based earcons) improve navigation performance in advanced auditory menus. *Human factors: The journal of human factors and ergonomics society*, 2012.

[127] B. N. Walker, A. Nance, and J. Lindsay. Spearcons: Speech-based earcons improve navigation performance in auditory menus. *Proceedings of ICAD '06*, pages 63–68, 2006.

[128] H. Wallach, E. B. Newman, and M. R. Rosenzweig. The precedence effect in sound localization (tutorial reprint). *Journal of the audio engineering society*, 21(10):817–826, 1973.

[129] J. Wang, S. Zhai, and J. Canny. SHRIMP - solving collision and out of vocabulary problems in mobile predictive input with motion gesture. In *Proceedings of CHI '10*, pages 15–24, 2010.

[130] M. Watson and P. Sanderson. Sonification helps eyes-free respiratory monitoring and task timesharing. *Human factors*, 46:497–517, 2004.

[131] D. Wigdor and R. Balakrishnan. TiltText: using tilt for text input to mobile phones. In *Proceedings of UIST*, pages 81–90. ACM, 2003.

[132] J. Williamson, R. Murray-Smith, and S. Hughes. Shoogle: excitatory multimodal interaction on mobile devices. In *Proceedings of CHI '07*, pages 121–124, New York, NY, USA, 2007. ACM.

[133] F. Winberg and J. Bowers. Assembling the senses: towards the design of cooperative interfaces for visually impaired users. In *Proceedings of CSCW '04*, volume 332-341, 2004.

[134] J. O. Wobbrock, M. R. Morris, and A. D. Wilson. User defined gestures for surface computing. In *Proceedings of CHI '09*, pages 1083–1092, 2009.

[135] M. Wozniewski, Z. Settel, and J. R. Cooperstock. A paradigm for physical interaction with sound in 3-d audio space. In *Proceedings of International computer music conference*, 2006.

[136] B. Yi, X. Cao, M. Fjeld, and S. Zhao. Exploring user motivations for eyes-free interaction on mobile devices. In *Proceedings of CHI '12*, pages 2789–2792, New York, NY, USA, 2012. ACM.

[137] M. Yin and S. Zhai. The benefits of augmenting telephone voice menu navigation with visual browsing and search. In *Proceedings of CHI '06*, pages 319–328, 2006.

[138] S. Zhao, P. Dragicevic, M. Chignell, R. Balakrishnan, and P. Baudisch. Earpod: Eyes-free menu selection using touch input and reactive audio feedback. In *Proceedings of CHI '07*, pages 1395–1404. ACM, 2007.

Aalto-DD 22/2013

ISBN 978-952-60-5002-7
ISBN 978-952-60-5003-4 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934
ISSN 1799-4942 (pdf)

Aalto University
School of Science
Department of Media Technology
www.aalto.fi

BUSINESS + ECONOMY

ART + DESIGN + ARCHITECTURE

SCIENCE + TECHNOLOGY

CROSSOVER

DOCTORAL DISSERTATIONS