Department of Media Technology

Enabling technologies for audio augmented reality systems

Hannes Gamper

Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 39/2014

Enabling technologies for audio augmented reality systems

Hannes Gamper

A doctoral dissertation completed for the degree of Doctor of Philosophy to be defended, with the permission of Aalto University School of Science, at a public examination held at lecture hall AS1 of the school on 2 May 2014, at 12 o'clock noon.

Aalto University
School of Science
Department of Media Technology

Supervising professor
Prof. Lauri Savioja

Thesis advisors
Assoc. Prof. Tapio Lokki
PhD Kai Puolamäki

Preliminary examiners
Assoc. Prof. Bruce N. Walker, Georgia Institute of Technology, Atlanta, United States of America

Assoc. Prof. Craig Jin, The University of Sydney, Sydney, Australia

Opponent
Dr Brian FG Katz, LIMSI-CNRS, Université Paris Sud, Orsay, France

Aalto University publication series
DOCTORAL DISSERTATIONS 39/2014

ISBN 978-952-60-5621-0
ISBN 978-952-60-5622-7 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-5622-7

Unigrafia Oy
Helsinki 2014

Finland

Publication orders (printed book):
The dissertation is available at https://aaltodoc.aalto.fi/

441 697
Printed matter

Aalto University

Abstract

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

Author
Hannes Gamper

Name of the doctoral dissertation
Enabling technologies for audio augmented reality systems

Publisher School of Science

Unit Department of Media Technology

Series Aalto University publication series DOCTORAL DISSERTATIONS 39/2014

Field of research Media Technology

Manuscript submitted 16 December 2013 **Date of the defence** 2 May 2014

Permission to publish granted (date) 20 March 2014 **Language** English

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

Abstract

Audio augmented reality (AAR) refers to technology that embeds computer-generated auditory content into a user's real acoustic environment. An AAR system has specific requirements that set it apart from regular human--computer interfaces: an audio playback system to allow the simultaneous perception of real and virtual sounds; motion tracking to enable interactivity and location-awareness; the design and implementation of auditory display to deliver AAR content; and spatial rendering to display spatialised AAR content. This thesis presents a series of studies on enabling technologies to meet these requirements.

A binaural headset with integrated microphones is assumed as the audio playback system, as it allows mobility and precise control over the ear input signals. Here, user position and orientation tracking methods are proposed that rely on speech signals recorded at the binaural headset microphones. To evaluate the proposed methods, the head orientations and positions of three conferees engaged in a discussion were tracked. The binaural microphones improved tracking performance substantially. The proposed methods are applicable to acoustic tracking with other forms of user-worn microphones.

Results from a listening test investigating the effect of auditory display parameters on user performance are reported. The parameters studied were derived from the design choices to be made when implementing auditory display. The results indicate that users are able to detect a sound sample among distractors and estimate sample numerosity accurately with both speech and non-speech audio, if the samples are presented with adequate temporal separation. Whether or not samples were separated spatially had no effect on user performance. However, with spatially separated samples, users were able to detect a sample among distractors and simultaneously localise it. The results of this study are applicable to a variety of AAR applications that require conveying sample presence or numerosity.

Spatial rendering is commonly implemented by convolving virtual sounds with head-related transfer functions (HRTFs). Here, a framework is proposed that interpolates HRTFs measured at arbitrary directions and distances. The framework employs Delaunay triangulation to group HRTFs into subsets suitable for interpolation and barycentric coordinates as interpolation weights. The proposed interpolation framework allows the realtime rendering of virtual sources in the near-field via HRTFs measured at various distances.

Keywords Audio augmented reality, acoustic tracking, auditory display, HRTF interpolation

ISBN (printed) 978-952-60-5621-0 **ISBN (pdf)** 978-952-60-5622-7

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

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2014

Pages 168 **urn** http://urn.fi/URN:ISBN:978-952-60-5622-7

Preface

The research work for this thesis has been carried out at the Department of Media Technology, Aalto University, during 2010–2013, and during a research visit in 2012 at the Department of Computer Science, University of Canterbury. The work was supported by the Helsinki Graduate School in Computer Science and Engineering (HeCSE), the MIDE program of Aalto University, the Nokia Research Foundation, and Tekniikan edistämissäätiö.

I thank my supervisor, Prof. Lauri Savioja, and my thesis advisors, Prof. Tapio Lokki and PhD Kai Puolamäki, for their support and advice, and for always having an open door and ear. I want to express my gratitude to Prof. Mark Billinghurst for hosting me at the Human Interface Technology Laboratory, and for collaborating with me and PhD Christina Dicke on a publication included in this thesis.

I thank the pre-examiners of this thesis, Assoc. Prof. Craig Jin and Assoc. Prof. Bruce N. Walker, for their invaluable feedback and comments that helped improve the thesis.

I thank my colleagues at the department, in particular the Virtual Acoustics team—Aki, Alex, Antti, Henna, Jonathan, Jukka P., Jukka S., Philip, Raine, Robert, Sampo, and Samuel—as well as the support staff for making this such a great place to work. Special thanks go to Dr Sakari Tervo for inspiring discussions, the collaboration in publications included in this thesis, and for proof-reading the thesis.

Finally, I thank my friends, my family, and especially Mandi for love and support throughout the years, and for reminding me that there is a life outside work.

Seattle, April 3, 2014,

Hannes Gamper

Contents

Preface **7**

Contents **9**

List of Publications **13**

Author's Contribution **15**

List of acronyms **19**

List of symbols **21**

1. Introduction **23**

1.1 Motivation 23

1.2 Scope of the thesis 24

1.3 Organisation of the thesis 25

2. Theoretical foundation **27**

2.1 Augmented reality 27

2.2 Audio augmented reality 28

2.3 Implementing an audio augmented reality system 29

2.4 Spatial hearing 31

2.4.1 Geometric definitions 31

2.4.2 Perception of lateral angle: Interaural cues 33

2.4.3 Perception of polar angle: Spectral cues 34

2.4.4 Perception of distance 35

2.4.5 Head-related transfer functions 36

2.4.6 Dynamic cues 36

2.4.7 Multi-modal cues 37

2.4.8 Properties and limitations of human spatial hearing . . 37

2.5 Spatial rendering 38

2.5.1 Playback systems 38
2.5.2 Rendering lateral angle: Interaural cues 40
2.5.3 Rendering polar angle: Spectral cues 41
2.5.4 Rendering distance and reverberation 42
2.5.5 Rendering using head-related transfer functions 43
2.5.6 Rendering dynamic cues 44
2.5.7 Properties and limitations of spatial rendering 45

3. Motion tracking **47**
3.1 Tracking techniques and systems 48
3.2 Acoustic tracking with particle filtering 50
3.2.1 Likelihood function 51
3.2.2 Particle filtering 52
3.3 Tracking speaker position 53
3.3.1 Voice activity detection 53
3.3.2 Time-delay estimation and likelihood function 54
3.3.3 Listener importance function 55
3.3.4 Particle filtering 56
3.4 Tracking listener orientation 57
3.4.1 Voice activity detection 57
3.4.2 Time-delay estimation and likelihood function 58
3.4.3 Particle filtering 59
3.5 Experimental setup 59
3.6 Results 61
3.6.1 Speaker location tracking 61
3.6.2 Orientation tracking 62
3.7 Discussion 65

4. Sound sample detection and numerosity estimation **67**
4.1 Related work 68
4.2 Experimental design and procedure 70
4.2.1 Test conditions 71
4.2.2 Apparatus and sound samples 72
4.2.3 Test procedure 72
4.3 Results 72
4.3.1 Task I: detect the <key> sample 73
4.3.2 Task II: estimate the <key> sample numerosity 74
4.4 Discussion 76

5. Rendering virtual sources **77**

5.1 Head-related transfer function interpolation 78

5.1.1 Subset selection 79

5.1.2 Calculation of interpolation weights 81

5.1.3 Interpolation in azimuth, elevation, and distance 81

5.2 Proposed approach 82

5.2.1 Triangulation of measurement points 83

5.2.2 Calculation of interpolation weights 83

5.2.3 Selecting a subset for interpolation 86

5.3 Experimental evaluation 87

5.4 Discussion 90

6. Summary **91**

6.1 Main results 91

6.2 Future work 92

Bibliography **95**

Publications **115**

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 H. Gamper, S. Tervo and T. Lokki. Head orientation tracking using binaural headset microphones. In *Proc. Int. Conv. Audio Engineering Society*, New York, NY, USA, paper number 8538, October 2011.

II H. Gamper, S. Tervo and T. Lokki. Speaker tracking for teleconferencing via binaural headset microphones. In *Proc. Int. Workshop on Acoustic Signal Enhancement (IWAENC)*, Aachen, Germany, 4 pages (online proceedings), September 2012.

III H. Gamper, C. Dicke, M. Billinghurst and K. Puolamäki. Sound sample detection and numerosity estimation using auditory display. *ACM Transactions on Applied Perception*, Vol. 10(1), pages 1–18, DOI:`http://dx.doi.org/10.1145/2422105.2422109`, February 2013.

IV H. Gamper. Selection and interpolation of head-related transfer functions for rendering moving virtual sound sources. In *Proc. Int. Conf. Digital Audio Effects (DAFx)*, Maynooth, Ireland, 7 pages (online proceedings), September 2013.

V H. Gamper. Head-related transfer function interpolation in azimuth, elevation, and distance. *J. Acoust. Soc. America*, 134(6), pages EL547–EL554, December 2013.

Author's Contribution

Publication I: "Head orientation tracking using binaural headset microphones"

A head orientation tracking method employing user-worn, binaural headset microphones is proposed. Unlike previous approaches from the literature, the proposed method does not require anchor sources, and instead relies on the users' speech signals. In a case study, the head orientations of three users in a meeting scenario were tracked. The average root-mean square error (RMSE) of the proposed method is about 10 degrees.

The present author had the original idea and wrote about 80% of the article. The development of the tracking algorithm and the experimental evaluation were done in collaboration with Dr. Sakari Tervo.

Publication II: "Speaker tracking for teleconferencing via binaural headset microphones"

The article proposes a position tracking algorithm employing user-worn, binaural headset microphones in combination with a reference microphone array. The tracking relies on speech signals recorded at the binaural microphones. Results of an experimental evaluation show the incorporation of binaural headset microphones into the tracking system to improve tracking accuracy substantially. The average root-mean square error (RMSE) of the proposed method is about 0.11 m.

The present author had the original idea and wrote about 80% of the article. The development of the tracking algorithm and the experimental evaluation were done in collaboration with Dr. Sakari Tervo.

Publication III: "Sound sample detection and numerosity estimation using auditory display"

This article investigates the effect of various auditory display design parameters on user performance in two basic tasks adapted from information visualisation, i.e., the detection of a sample among distractors, and the estimation of sample numerosity. Sets of sound samples were presented to test participants in a listening test. In the test, the stimulus onset asynchrony (SOA) of the samples had a substantial effect on user performance in both tasks, in contrast to the sound type and spatial quality of the samples, which had a minor effect. The results suggest that diotic or indeed monophonic playback with appropriately chosen SOA may be sufficient in practical applications requiring users to detect a sample or estimate sample numerosity. However, if spatial information was present in the samples, the test subjects were able to simultaneously detect and localise a sample with reasonable accuracy.

The development and implementation of the user study was a joint effort of Dr. Kai Puolamäki, Dr. Christina Dicke, and the present author. The present author performed the data analysis with the help of Dr. Kai Puolamäki, and wrote about 90% of the article.

Publication IV: "Selection and interpolation of head-related transfer functions for rendering moving virtual sound sources"

The article studies the selection of head-related transfer function (HRTF) measurements on the surface of a sphere for interpolation, and the calculation of linear interpolation weights. An HRTF interpolation framework is proposed based on a method for subset selection and interpolation weight calculation that is independent of the HRTF measurement grid layout. The proposed method relies on Delaunay triangulation to group HRTF measurements into non-overlapping triplets, and uses vector base amplitude panning (VBAP) gains for interpolation. An experimental evaluation shows the proposed framework to be robust against grid irregularities and to be suitable for rendering dynamic virtual sound sources.

The present author is the sole author of this article.

Publication V: "Head-related transfer function interpolation in azimuth, elevation, and distance"

The article extends the head-related transfer function (HRTF) subset selection and interpolation weight calculation framework presented in publication IV for HRTF measurements obtained at various distances. The proposed framework relies on Delaunay triangulation to group HRTFs into subsets for interpolation, barycentric coordinates as linear interpolation weights, and a fast search algorithm to find a suitable subset for interpolation. The proposed framework is robust with respect to grid irregularities and computationally efficient. An experimental evaluation shows good accordance between measured and interpolated HRTFs.

The present author is the sole author of this article.

List of acronyms

2-D	two-dimensional
3-D	three-dimensional
AAR	audio augmented reality
AR	augmented reality
BRIR	binaural room impulse response
BRTF	bone-related transfer function
FFT	Fast Fourier Transform
GPS	Global Positioning System
HRIR	head-related impulse response
HRTF	head-related transfer function
IHL	inside-the-head locatedness
IIR	infinite impulse response
ILD	interaural level difference
IR	infrared
ITD	interaural time difference
MIDI	musical instrument digital interface
MLE	maximum likelihood estimation
PDF	probability density function
RMSE	root-mean square error
SNR	signal-to-noise ratio
SOA	stimulus onset asynchrony
SRM	spatial release from masking
TDOA	time-difference of arrival
TOA	time of arrival
VBAP	vector base amplitude panning
VR	virtual reality
WFS	wave field synthesis
WLAN	Wireless Local Area Network

List of symbols

θ	elevation angle
φ	azimuth angle
r	radius
γ	lateral angle
δ	polar rotation angle
Δ	difference
t	time of arrival (TOA)
τ	time-difference of arrival (TDOA)
a	effective head radius
d	distance
f_s	audio sampling rate
c	speed of sound
$\mathbf{r}$	receiver position
$\mathbf{s}$	source/speaker position
$\|\cdot\|$	Euclidean norm
$x(t)$	time-domain signal
$X(f)$	frequency-domain signal
$(\cdot)^*$	complex conjugate
$\arg\max$	argument of the maximum
σ	standard deviation
$\hat{\cdot}$	estimate
w	particle weight
$\mathbf{p}$	particle position
$p(\cdot\|\cdot)$	likelihood function

1. Introduction

Audio augmented reality (AAR) is a technology that aims to embed virtual auditory content into the real environment of a user. This thesis studies some of the challenges involved in implementing an AAR system, and presents possible approaches to resolve them.

1.1 Motivation

Nearly five decades after the first augmented reality (AR) application was presented by Sutherland (1968), the technology is still at an early stage in its development (Nicholson, 2013), and has only recently reached the general public in the form of advertising, augmented sports broadcasting (Olaizola et al., 2006) and mobile AR browsers, including Wikitude[1], Layar[2], and Junaio[3]. While the above examples are mainly based on visual display of augmented content, relatively few applications that run outside laboratory settings provide auditory augmentation. An example of such an application is the mobile AAR browser Toozla[4]. Possible reasons for the slow adoption of AAR include a general trend in human–computer interaction research to give prevalence to the human vision over other senses (Cohen and Wenzel, 1995), a lack of AR authoring tools supporting audio, and perhaps uncertainty among AR application designers regarding the benefits and requirements of AAR.

This thesis summarises a series of studies on enabling technologies for AAR, from motion tracking to auditory display and spatial rendering. These studies helped to identify the challenges and requirements of an AAR system, and resulted in some novel approaches to overcome them.

[1] `www.wikitude.com`
[2] `www.layar.com`
[3] `www.junaio.com`
[4] `www.toozla.com`

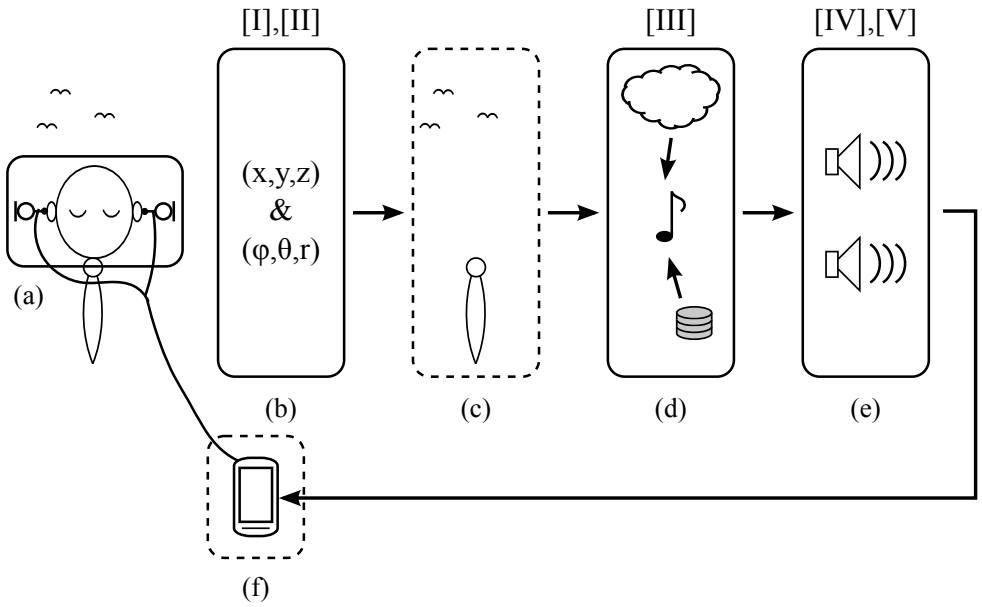

Figure 1.1. AAR system overview: (a) audio playback setup; (b) motion tracking; (c) context extraction; (d) audio encoding; (e) spatial rendering; (f) user interface. The parts studied in this thesis are marked along with Roman numerals indicating the respective publications.

1.2 Scope of the thesis

The research leading to this thesis was motivated by issues and challenges arising when designing and implementing an AAR system, from the choice of playback setup and tracking technology to the design of auditory display and the spatial rendering framework. While this thesis does by no means strive to address all topics relevant to the research area, it does highlight some of the problems and potential pitfalls an AAR system or application designer might encounter, and discusses both previously presented and novel approaches to resolve them. Figure 1.1 shows a block diagram of an AAR system. The basic building blocks of the AAR system are (a) the audio playback setup, (b) motion tracking, (c) context extraction, (d) audio encoding, and (e) spatial rendering.

In this thesis, a binaural headset with integrated microphones is assumed as the audio playback setup (see Fig. 1.1a). It allows the user to perceive both real and virtual environments simultaneously as *augmented reality*.

Motion tracking (see Fig. 1.1b) determines the position and orientation of the user and is required in AR systems to register the augmentation layer with the real environment. A variety of motion tracking methods and systems have been proposed previously. Here, a method is presented to extract position and orientation information from the signals of the user-worn headset microphones.

Context extraction (see Fig. 1.1c) describes the process of determining the

context the user is in, based on features including user location (Liao et al., 2007), and presence or absence of people or objects in the environment (Ajanki et al., 2011). Context-awareness allows an AR application to deliver virtual content that is relevant or interesting for the given situation (Ajanki et al., 2011), and thus *augments* the perception of the real environment. However, the extraction and interpretation of context is highly application-specific, and not part of the present work.

Audio encoding (see Fig. 1.1d) is the process of making virtual content audible. Building on related research on enabling technologies for auditory display, the work presented here investigates the effect of various display design parameters on user performance. The parameters studied include the sound type of the audio samples used to display information and their arrangement in time and space. The user performance is evaluated in two basic tasks adapted from information visualisation: detecting a sample among distractors, and estimating sample numerosity. Due to the general nature of the tasks, the results of the study have potential implications for a variety of practical applications, and may inform the choice of enabling technology for auditory display in an AAR setup.

Spatial rendering (see Fig. 1.1e) is the process of generating ear input signals that evoke the perception of a virtual sound source emanating from a specific direction or position in space. In AAR, virtual content is displayed via spatialised virtual sound sources as an overlay onto the real acoustic environment. The rendering process encodes measured and/or modelled localisation cues into the sound signal of a virtual source. Here, a spatial rendering framework is proposed that produces virtual sources with high fidelity and is not tied to a specific database of localisation cues, unlike previously proposed approaches.

Optionally, an AAR system may also require interfaces to support user interaction. The study of such interfaces is closely related to human–computer interaction research, and is outside the scope of this thesis.

1.3 Organisation of the thesis

Chapter 2 presents the theoretical foundation of this thesis. The properties of human spatial hearing are discussed, as well as the application of those properties for rendering spatialised virtual content. Chapter 3 introduces the motion tracking algorithms employing the microphone signals of the user-worn AAR headset, as proposed in Publications I and II. Chapter 4 discusses the use of auditory display to convey sample presence or numerosity, and presents

results from a listening test reported in Publication III. A rendering framework for displaying spatialised virtual audio content, published in Publications IV and V, is introduced in Chapter 5. Chapter 6 summarises and concludes the thesis.

2. Theoretical foundation

This chapter gives an overview of the theoretical context of this thesis. First, definitions of augmented reality (AR) and audio augmented reality (AAR), as used in this thesis, are presented. Then, the requirements for implementing an AAR system are briefly discussed. Finally, a short review of the perception and generation of spatial sound is given, as these form the basis of AAR.

2.1 Augmented reality

AR aims at enhancing the sensory perception of the real world by embedding computer-generated, virtual stimuli or information into the user's environment (Azuma, 1997; Rozier et al., 2000). Azuma et al. (2001) define AR as a variation of virtual reality (VR), with the following properties:

- combines real and virtual objects in a real environment;
- runs interactively, and in real time;
- registers (aligns) real and virtual objects with each other.

An alternative interpretation places AR between real and virtual environments on a reality–virtuality continuum (Milgram et al., 1995), as it combines real and virtual elements.

The first AR application dates back to 1968, when Sutherland presented a see-through head-mounted display that showed three-dimensional (3-D) information with a "kinetic depth effect" (Sutherland, 1968): The perspective of the displayed information changes in accordance with head movements of the viewer, to give the illusion of a 3-D object. The possibility of embedding virtual content into the perception of the real environment through AR has since found use in a variety of applications, including television broadcasting (Olaizola et al., 2006), medical displays (Azuma, 1997; Sielhorst et al., 2008), and industrial applications (Regenbrecht et al., 2005; Pentenrieder et al., 2007).

With the advent of powerful portable computers and mobile phones, mobile AR applications emerged, allowing the augmentation of the real world outside laboratory settings (Feiner et al., 1997; Starner et al., 1997; Henrysson and Ollila, 2004; Ajanki et al., 2011).

2.2 Audio augmented reality

Although many AR applications rely mostly on visual augmentation of reality, research on taking advantage of sensory modalities other than vision is growing, not least to make AR accessible to the blind and visually impaired. AAR can be defined analogously to (visual) AR as a combination of real and virtual auditory objects in a real environment (Warusfel and Eckel, 2004). Audio forms an interesting alternative to vision as a display modality in AR, for a number of reasons. The goal of AR is to enhance, rather than replace, reality. An AR system must therefore support the simultaneous perception of the real environment and the virtual overlay. This is especially important in a mobile context, where the user should be continuously aware of the surroundings (McGookin and Brewster, 2004b). Given the user's limited field of view, using a graphical interface can be challenging in situations where the user is engaged in a visually demanding task, such as walking or driving. These limitations can be overcome with a non-graphical display. An example of a non-graphical display is *auditory display*, defined as "the use of sound to communicate information about the state of a computing device to a user" (McGookin and Brewster, 2004b). A key advantage of auditory over graphical display is that it does not require a stable line of sight and is not limited to a "field of view". Therefore, in AAR, information can be presented to the user via auditory display regardless of the user's head orientation. Furthermore, channeling information to the ears reduces the visual and cognitive load and frees the user's eyes to observe the environment (Peres et al., 2008). A combination of visual and auditory information display can be beneficial for multimodal tasks (Hornof et al., 2010) or to improve the usability of a device with a small visual display (Brewster, 2002). While the sense of vision outperforms the auditory system in terms of its spatial resolution (Behringer et al., 1999), the auditory system has a higher temporal resolution and may react faster to stimuli than the visual system (Nees and Walker, 2009). In an alerting or monitoring task, the auditory system is able to rapidly detect unexpected sounds, while ignoring expected ones (Shinn-Cunningham et al., 1997), and to attend multiple audio streams in parallel (Bregman, 1990). A listener can focus on a particular speaker among a

group of concurring speakers, a phenomenon referred to as the "cocktail party effect" (Cherry, 1953). Based on the properties of the human auditory system, researchers have identified use cases for auditory display in a variety of AR scenarios, including telecommunication (Dalenbäck et al., 1996; Beracoechea et al., 2008), navigation (Loomis et al., 1998; Sundareswaran et al., 2003), tour guiding (Bederson, 1995; Zimmermann and Lorenz, 2008), context-aware computing (Mynatt et al., 1998; Sawhney and Schmandt, 2000) and device diagnostics and maintenance (Behringer et al., 1999).

2.3 Implementing an audio augmented reality system

Table 2.1 lists examples of AAR systems and their components. Despite the variety of application areas, the systems share the basic building blocks depicted in Fig. 1.1. All systems require an audio playback setup and some form of audio encoding, to display audible content to the user. For the playback setup, most systems rely on user-worn headphones, as they are both cheap and portable. The form of audio encoding employed is somewhat application specific. Guiding and navigation systems benefit from synthesised or pre-recorded speech output, to provide explicit information to the user. Non-speech sounds, on the other hand, may be required to alert the user, provide background information or awareness, or communicate other non-verbal cues, for instance the spatial location of an object or place.

Motion tracking is a part of all but one system. Knowing the position of the user allows the AAR system to provide location-dependent information. In many systems, the user context is inferred simply from user location. Furthermore, location-awareness enables implicit user interaction: The displayed auditory content changes as the user moves. For many systems this type of passive user interaction is sufficient or even preferred (McGookin and Brewster, 2012), and no dedicated user interface is required.

Most AAR systems listed in Table 2.1 employ spatial rendering to display virtual auditory content at arbitrary directions or locations. Spatial rendering extends the auditory display space beyond the physical boundaries of the playback setup's transducers, creating what may be referred to as *virtual auditory display* (Shilling and Shinn-Cunningham, 2002). In the following, a short overview of the human ability to perceive and localise sound is given, followed by a brief review of spatial rendering.

AAR system	Application	Playback setup	Motion tracking	Content extraction	Audio encoding	Spatial rendering	Interaction
Mobile spatial audio communication system (Kan et al., 2004)	telecommunication	headphones	GPS	-	live recording	HRTF filtering	passive
Virtual acoustic opening (Beracoechea et al., 2008)	telecommunication	loudspeaker array	a priori knowledge	-	live recording	WFS	passive
Personal guidance system (Loomis et al., 1998)	navigation	headphones	GPS + compass	location	synthesised speech	HRTF filtering	keypad
(Sundareswaran et al., 2003)	navigation	headphones	GPS + magnetometer	location	auditory icons	HRTF filtering	speech, buttons
SWAN (Wilson et al., 2007)	navigation	bonephones	GPS + inertial	location, history	auditory icons, earcons, spearcons	BRTF* filtering	tactile
Automated Tour Guide (Bederson, 1995)	museum guide	headphones	IR badges	location	pre-recorded	-	passive
LISTEN (Zimmermann and Lorenz, 2008)	museum guide	headphones	radio-frequency beacon	location, history	pre-recorded, auditory icons	HRTF filtering	passive
Audio Aura (Mynatt et al., 1998)	messaging	headphones	IR badges	location	speech, auditory icons, earcons	-	passive
Nomadic Radio (Sawhney and Schmandt, 2000)	messaging	wearable speakers	-	message priority, usage level, acoustic environment	ambient, pre-recorded, auditory icons, synthesised speech	HRTF filtering	speech, buttons
Device Diagnostics Systems (Behringer et al., 1999)	maintenance	loudspeakers, headphones	fiducial	-	pre-recorded	HRTF filtering	speech
PULSE (McGookin and Brewster, 2012)	social computing	headphones	iPhone geolocation	location	synthesised speech, auditory icons	HRTF filtering	passive

*bone-related transfer function (BRTF)

Table 2.1. Overview of AAR systems and their components (cf. Fig. 1.1).

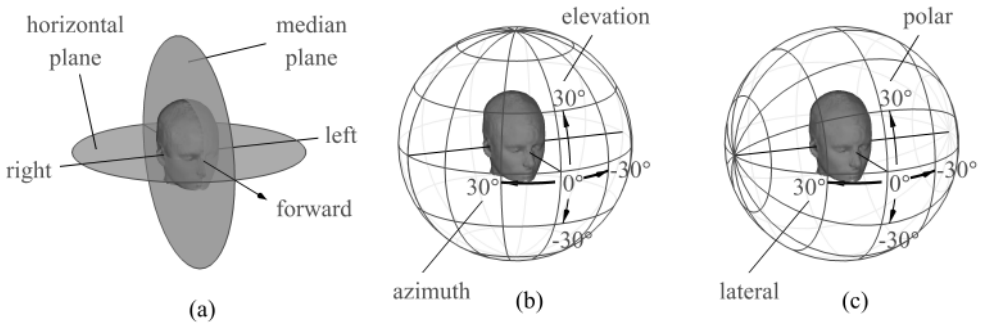

Figure 2.1. (a) Head-related coordinate system; (b) vertical-polar, and (c) horizontal-polar coordinate system. The head model is taken from the EEGLAB toolbox (Delorme and Makeig, 2004).

2.4 Spatial hearing

To generate and render virtual auditory events embedded into a real physical environment, the properties of real auditory events as well as their perception by the human auditory system need to be taken into account. *Hearing* can be defined as the perception of auditory events that occur at a certain time and place. Therefore, human hearing is inherently spatial (Blauert, 1996). Through *localisation*, the auditory system relates attributes of the sound reaching the ears to the location of an auditory event. In the following, these sound attributes and their role for determining the position of an auditory event are briefly reviewed.

2.4.1 Geometric definitions

In this thesis, geometric relations are described in the head-related coordinate system described by Blauert (1996), unless otherwise stated. The coordinate system is depicted in Fig. 2.1. The following geometric definitions are used throughout this thesis:

Origin The origin of the coordinate system lies halfway between the ear entrances.

Horizontal plane The plane through the origin intersecting the ear entrances and eye sockets, dividing the space into upper and lower hemisphere (see Fig. 2.1a).

Median plane The plane orthogonal to the horizontal plane and halfway between the eye sockets, dividing the space into left and right hemisphere (see Fig. 2.1a).

Head symmetry Symmetry of the head about the median plane (Blauert, 1996).

Elevation The angle, $-90 \leq \theta \leq 90$, measured between the horizontal plane and a ray from the origin to a 3-D location; an elevation of -90 degrees lies below the head, an elevation of 90 degrees lies above the head (see Fig. 2.1b).

Azimuth The angle, $-180 \leq \varphi \leq 180$, measured between the median plane and the projection of the ray from the origin to a 3-D location onto the horizontal plane; an azimuth of -90 degrees lies to the left, an azimuth of 90 degrees lies to the right, and an azimuth of ± 180 degrees lies behind the head (see Fig. 2.1b).

Radius The distance, r, from the origin.

Vertical-polar coordinate system Describes 3-D location in terms of azimuth, φ, elevation, θ, and radius, r (Macpherson and Middlebrooks, 2002) (see Fig. 2.1b).

Horizontal-polar coordinate system Describes 3-D location in terms of lateral angle, γ, polar angle, δ, and radius, r (Macpherson and Middlebrooks, 2002) (see Fig. 2.1c).

Lateral angle The angle, $-90 \leq \gamma \leq 90$, measured between the median plane and a ray from the origin (Algazi et al., 2001b); a lateral angle of -90 degrees lies to the left, a lateral angle of 90 degrees to the right of the head (see Fig. 2.1c).

Polar angle The polar rotation angle, $-180 \leq \delta \leq 180$, in the horizontal-polar coordinate system (Algazi et al., 2001b); a polar angle of -90 degrees lies below, a polar angle of 90 degrees above, and a polar angle of ± 180 degrees behind the head (see Fig. 2.1c).

Near-field The region about 1 m or less away from a listener's head (Kan et al., 2009).

Far-field The region further than about 1 m away from a listener's head (Kan et al., 2009).

Vertical-polar coordinates, φ, θ, can be converted to horizontal-polar coordinates, γ, δ, via (Morimoto and Aokata, 1984)

$$\gamma = \arcsin\left(\sin\varphi\cos\theta\right), \tag{2.1}$$

$$\delta = \begin{cases} \delta' & \text{if } |\varphi| < \frac{\pi}{2}, \\ \pi - \delta' & \text{else,} \end{cases} \tag{2.2}$$

where

$$\delta' = \arcsin\left(\frac{\sin\theta}{\sqrt{\sin^2\theta + \cos^2\varphi\cos^2\theta}}\right). \tag{2.3}$$

2.4.2 Perception of lateral angle: Interaural cues

Early experiments on human localisation have demonstrated the ability of humans to determine the direction of pure tones based on differences between the signals reaching the left and right ear (Rayleigh, 1907; Macpherson and Middlebrooks, 2002). For low-frequency pure tones, the auditory system primarily evaluates phase differences between the ear signals to determine the lateral angle of a sound source. For frequencies above 500 Hz, the lateral angle of a source can be inferred from level differences between the ear signals. As these localisation cues stem from differences between the ear signals, they are referred to as *interaural cues* (Blauert, 1996).

Real auditory events carry interaural cues due to the physics underlying the propagation of sound in air. Sound emanating from a sound source that is small compared to the wavelength of the sound propagates in spherical longitudinal waves (Rossing and Fletcher, 2004). If the sound source is positioned to the left or to the right of a listener, the propagation paths from the source to each ear of the listener differ in length. Therefore, the wave front first reaches the ipsilateral ear (i.e., the ear oriented towards the source), and then the contralateral ear (i.e., the ear oriented away from the source). The signal reaching the contralateral ear is subject to a delay proportional to the difference in path lengths. This delay is referred to as interaural time difference (ITD). The ITD changes as a function of the source's lateral angle, and can therefore be evaluated by the auditory system as a cue for the lateral direction of the sound source. For pure tones with a frequency up to 1.5–1.6 kHz, the ITD can be derived from the phase difference between the signals at the ipsilateral and the contralateral ear. At higher frequencies, the wavelength is shorter than the distance between the ears, i.e., shorter than about 20 cm (Blauert, 1996). Therefore, the wave may cycle from the moment it reaches the ipsilateral ear to the moment it reaches the contralateral ear. The resulting phase difference

between the ear signals is ambiguous, hence the ITD can not be inferred from it. It should be noted that for complex high-frequency sounds the auditory system is able to extract ITD information from the onsets and envelope of the ear signals (Macpherson and Middlebrooks, 2002).

The dominant cue for determining the lateral angle of high-frequency sounds is the interaural level difference (ILD) between the ear signals (Rayleigh, 1907; Macpherson and Middlebrooks, 2002). The ILD is a result of the listener's head causing acoustic shadowing, thus reducing the signal level at the contralateral ear (Blauert, 1996). Towards low frequencies with wavelengths larger than the head size, the head becomes acoustically transparent and the ILD diminishes.

The relative importance of the interaural cues for determining the lateral angle of a source is explained by the *Duplex theory* (Rayleigh, 1907; Macpherson and Middlebrooks, 2002): The auditory system weights ITD cues strongly in the low-frequency region and ILD cues strongly in the high-frequency region.

2.4.3 Perception of polar angle: Spectral cues

The auditory system uses interaural cues described in Section 2.4.2 to determine the perceived lateral angle of a sound source. However, for sound sources in the far-field positioned on the median plane or a cone centred on the axis connecting the ears (i.e., a *cone of confusion*), and assuming free-field conditions and a symmetrical head without torso, these cues are invariant (Hebrank and Wright, 1974; Shinn-Cunningham et al., 2000). Nevertheless, the auditory system is able to extract elevation cues from sufficiently long or repeated broadband signals of a source on the median plane (Blauert, 1996). These cues are monaural, as the ear signals they are extracted from are identical. It has been shown that the auditory system is not able to interpret monaural temporal cues, and that elevation perception is instead based on monaural spectral cues (Hebrank and Wright, 1974; Wightman and Kistler, 1997). Studies have shown that the impression of source elevation can be created by applying a notch (Bloom, 1977) or peak (Blauert, 1996) with elevation-dependent centre frequency to the signal spectrum. In the case of real auditory events, elevation-dependent spectral peaks and notches are caused by pinna and torso reflections (Zotkin et al., 2004; Takemoto et al., 2012). The combination of these spectral peaks and notches is believed to serve as an elevation cue (Wightman and Kistler, 1997; Zotkin et al., 2004; Takemoto et al., 2012).

Experiments by Macpherson and Middlebrooks suggest that monaural spectral cues have little or no importance for lateral angle perception (Macpherson and Middlebrooks, 2002). Interaural spectral cues, that is, frequency-dependent

ILD patterns, have been suggested as elevation cues (Duda, 1997) and lateral angle cues (Macpherson and Middlebrooks, 2002), but their role seems to be minor (Macpherson and Middlebrooks, 2002; Jin et al., 2004).

For the human auditory system to be able to extract spectral cues from the ear signals, some prerequisites should be met. Firstly, although it has been suggested that monaural spectral features exist below 3 kHz (Algazi et al., 2001b), the source signal should have spectral content above 5 kHz (Wightman and Kistler, 1997). Secondly, the auditory system should have prior knowledge of the source signal, that is, it should be familiar with the source sound (Blauert, 1996). Thirdly, and perhaps most importantly, the auditory system needs prior knowledge of the way the spectral cues change as a function of the source direction. These spectral patterns are discussed in Section 2.4.5.

2.4.4 Perception of distance

Determining the distance of a sound source is quite a challenging task for the human auditory system (Zahorik et al., 2005). Distance perception is based on a variety of factors. A straightforward cue to judge the distance of a sound source is the sound intensity: The sound is attenuated as it propagates, hence the intensity increases as the sound source approaches the listener. For moving sources, the rate at which the sound intensity changes can be used by listeners to judge source distance (Zahorik et al., 2005). An important distance cue for sources in reverberant environments is the ratio between direct and reverberant sound energy (Middlebrooks and Green, 1991; Zahorik et al., 2005): Close sound sources have a higher direct sound energy relative to the reverberant sound energy than further sources. For sources further than about 15 m from the listener, the high-frequency attenuation due to air absorption can serve as a distance cue (Zahorik et al., 2005). For sources in the near-field, it has been suggested that the ILD changes differently with the source position than the ITD (Shinn-Cunningham et al., 2000; Brungart, 2002). While the ITD is largely unaffected by source distance, the ILD for a lateral source increases with decreasing distance. The reasons for the ILD boost in the near-field are an increased effect of head shadowing and the fact that for a sound source approaching the head, the level of the ipsilateral ear signal increases faster than the level of the contralateral ear signal (Brungart, 2002). The faster increase of the ipsilateral signal level leads to an ILD boost at low frequencies that exceeds low-frequency ILDs found in the far-field.

An important non-acoustic cue for distance perception of an auditory event is the familiarity of the listener with the source signal (Blauert, 1996; Zahorik

et al., 2005). Human listeners can determine the distance of a live talker reasonably well (Middlebrooks and Green, 1991; Zahorik et al., 2005), but fail to determine the distance of unfamiliar sounds, unless reverberation is present allowing the listeners to judge the distance based on the direct-to-reverberant energy ratio (Brungart, 2002).

2.4.5 Head-related transfer functions

The localisation cues contained in the sound signal of a source in free field are a result of the filtering that sound undergoes when travelling from a sound source to the listener's ears due to shadowing and reflections from the listener's torso, head, and pinnae (Middlebrooks et al., 1989; Wightman and Kistler, 1989). Assuming this filtering to be linear and time-invariant, it can be described by an impulse response or a transfer function (Breebaart, 2013), the *head-related impulse response (HRIR)* or *head-related transfer function (HRTF)*, respectively. The HRTF can be defined as the relation of the sound pressure at a point inside the human ear canal to the sound pressure at the centre of the head in absence of the listener (Blauert, 1996). As the HRTFs are highly dependent on the lateral and polar angle of the sound source, they contain the lateral and polar localisation cues described in Sections 2.4.2 and 2.4.3. For sources in the near-field, HRTFs are distance-dependent (Brungart, 2002; Kan et al., 2009), and hence capture some of the distance cues mentioned in Section 2.4.4.

An important characteristic of HRTFs is that they are highly individual, due to differences in the geometric and acoustic properties of the torso, head, and pinnae between listeners (Wenzel et al., 1993). HRTFs can be measured by inserting probe microphones into the ear canals of a listener. Databases of HRTF measurements are publicly available online (Gardner and Martin, 1995; Algazi et al., 2001a; IRCAM, 2013). Measuring and analysing HRTFs is of ongoing research interest as it allows studying the acoustic cues responsible for human sound localisation. The usage of HRTFs for rendering spatialised audio is discussed in Section 2.5.5.

2.4.6 Dynamic cues

The localisation accuracy of the auditory system is best for sources straight ahead of the listener (Middlebrooks and Green, 1991; Blauert, 1996). To determine the position of a sound source, listeners tend to spontaneously move the head towards it to improve the localisation accuracy (Middlebrooks and Green, 1991; Blauert, 1996). This head movement results in a change of

the localisation cues encoded in the ear signals. The patterns with which the localisation cues change due to head movements constitute dynamic localisation cues (Blauert, 1996). People who are deaf on one ear can use these dynamic cues to better localise a sound source (Blauert, 1996). For listeners with normal hearing, dynamic cues seem to serve mostly for resolving localisation ambiguities, e.g., to determine whether a source is in front of or behind the listener (Blauert, 1996).

2.4.7 Multi-modal cues

Not all factors that affect the auditory perception are themselves strictly auditory (Slaney, 1998). To be able to extract the dynamic cues discussed in Section 2.4.6 from the ear signals, the listener must relate them to the head movements that caused them. The head movement is in turn inferred from the senses of vision and balance, and the position of the neck muscles. Therefore, dynamic cues can be considered multi-modal (Blauert, 1996). There are several other examples where the sense of vision affects auditory perception. Visual feedback has been shown to improve speech intelligibility in the presence of noise or competing speech (Bernstein and Grant, 2009). In the case of conflicting auditory and visual cues, the sense of vision may dominate the auditory perception. If the temporal changes of a visual object are synchronised to the changes of sound signal, the viewer might localise the sound source at the position of the visual object, even if the actual sound source is located at a different position (Yost, 1993; Blauert, 1996). This phenomenon, referred to as *visual capture*, can be experienced when watching a television programme or a ventriloquist: Sound synchronous to lip movements is heard as emanating from a person displayed on screen or the ventriloquist's puppet, even though the sound does not actually originate from there. The *McGurk effect* demonstrates how the visual perception of lip movements influences the auditory perception of speech sounds (Cohen and Massaro, 1990): A video of a person articulating /pa-pa/ combined with the speech sounds /na-na/ can result in the viewer hearing /ma-ma/.

2.4.8 Properties and limitations of human spatial hearing

The accuracy of human auditory localisation can be described in terms of the *localisation blur*, i.e., the minimum sound source displacement perceivable by 50% of listeners (Blauert, 1996). For sound sources straight ahead, listeners are able to detect lateral displacements as small as one degree. This is taken

as the maximum spatial resolution of the human hearing. The localisation blur increases with the source azimuth, reaching a maximum to either side of the listener. The localisation blur is higher in the vertical direction than in the horizontal direction. The minimum localisation blur for the vertical displacement or a source straight ahead of the listener is about four degrees for white noise, nine degrees for a talker familiar to the listener, and 17 degrees for unfamiliar speech (Blauert, 1996). For a source above or behind the listener, the localisation blur in vertical direction increases.

The localisation of a sound source in a reverberant environment is aided by the *precedence effect* (Litovsky and Godar, 2010): The auditory system emphasises localisation cues encoded in the sound reaching the ears on a direct path from the source, while de-emphasising cues stemming from reflections that incur a propagation delay relative to the direct sound.

Due to the "cocktail party effect" (Cherry, 1953; Blauert, 1996), the auditory system is able to employ a set of temporal, spectral, and spatial cues to follow a target speaker in the presence of competing sound sources (Yost, 1997).

An overview of the auditory system's performance in a variety of basic discrimination and identification tasks is given by Kidd et al. (2007).

2.5 Spatial rendering

When generating audio feedback in augmented reality, the ability to position auditory events is necessary to allow them to be overlaid over the real acoustic environment. Based on the understanding of human spatial hearing (see Section 2.4), it is possible to render a virtual sound source and control the way it is perceived by a listener. The process of rendering virtual sound in such a way that it evokes the same listening experience as a real sound source at a specific point in space is referred to as *auralisation* (Kleiner et al., 1993). Next, playback systems and rendering techniques for auralisation in AAR are discussed.

2.5.1 Playback systems

The rendering of spatial audio requires precise control over the signals reaching the listener's ears. Controlling the ear input signals of the left and the right ear independently allows to encode the spatial cues that evoke the perception and localisation of an auditory event. A playback system for spatial rendering has to support channel separation at the ears of the listener, to enable the

faithful delivery of these spatial cues. A playback system employing a pair of loudspeakers to control the ear input signals of a listener is referred to as *transaural stereo* (Cooper and Bauck, 1989). With a transaural loudspeaker setup, the channel separation required for spatial rendering is achieved by ensuring the signal of one loudspeaker reaches only one ear, via a crosstalk cancellation algorithm (Atal and Schroeder, 1966; Gardner, 1997). The crosstalk cancellation typically only works if the listener remains within a restricted area known as the "sweet spot" (Gardner, 1997). A major drawback of using loudspeakers for spatial rendering in an augmented reality system is that it does not support mobility of the user. A mobile variant of loudspeaker-based systems is the "Soundbeam Neckset" that comprises user-worn directional loudspeakers (Sawhney and Schmandt, 2000).

Headphone-based systems for spatial rendering provide the advantages that they are portable and have high channel separation, allowing precise control over the ear input signals (Shilling and Shinn-Cunningham, 2002). In an augmented reality setup, the use of headphones may be problematic due to the occlusion of the user's ear canals, which may deteriorate the perception of the real acoustic environment. Awareness of one's surroundings is especially important in mobile applications, to alert the user of potential dangers. To enhance the perception of ambient sounds when wearing headphones, Tappan (1964) proposed the use of "Nearphones", i.e., small loudspeakers worn near the ears. Bone-conductive headsets, or "bone-phones" (Walker and Lindsay, 2005), transmit sound to the cochlea by inducing vibrations directly to the skull, and thus do not occlude the ear entrances. Bone-phones have been successfully used to render spatialised audio (MacDonald et al., 2006) and "hear-through augmented reality" (Lindeman et al., 2007). Martin et al. (2009) propose the use of earphones equipped with acoustically transparent earpieces to enable hear-through augmented reality. "Mic-through augmented reality" (Lindeman et al., 2007), on the other hand, refers to the use of headphones with integrated microphones for spatial rendering. Playing back the microphone signals to the user mitigates the attenuation of ambient sounds due to the headphones occluding the ear entrances. Some commercially available noise-cancelling earphones employ "mic-through" technology to improve the perception of ambient sounds: Sennheiser[1] equips some of their noise-cancelling headphone models with "TalkThrough" technology (Gelhard and Grone, 2010), whereas Bose's "QuietComfort" earbuds[2] come with an "Aware" mode. An example

[1] `www.sennheiser.com`
[2] `www.bose.com/qc`

Figure 2.2. The ARA headset and mixer. See text for description.

of a headset specifically geared towards AAR applications is the "Intelligent Headset" by GN Store Nord[3].

A similar concept underlies the "ARA headset" (Härmä et al., 2004; Albrecht et al., 2011; Rämö and Välimäki, 2012) that was designed to enable the rendering of spatial audio for mobile AR applications. The ARA headset, shown in Fig. 2.2, consists of a pair of insert-earphones with integrated miniature microphones, and a mixer. The real acoustic environment is captured at the user's ear entrances via the microphones and played back through the earbuds. The microphone signals are equalised in the mixer to minimise the effect of the headset on the captured sounds (Albrecht et al., 2011), with the goal of making the headset acoustically transparent. The audio augmentation is implemented by playing back virtual sounds through the earbuds. Therefore, the ARA headset allows rendering virtual content overlaid onto reality, while maintaining high fidelity with respect to the perception of the real acoustic environment. The level of the microphone signals can be adjusted in the ARA mixer to either amplify or attenuate ambient sounds, allowing the user to crossfade between real and virtual content. In the remainder of this thesis, the audio AR system is assumed to rely on a headphone-based playback system such as the ARA headset and mixer.

2.5.2 Rendering lateral angle: Interaural cues

The process of rendering spatialised virtual audio via binaural headphone signals is referred to as *binaural synthesis* (Jot et al., 1995). In the context of augmented reality, the goal of binaural synthesis is to render a virtual sound source in such a way that it is perceived by the user as being embedded in the real acoustic environment. The degree of fidelity of the spatial rendering depends on a variety of factors, including the requirements of the AR application and the constraints of the AR system. A straightforward way to spatialise a monophonic input source via binaural synthesis is to encode basic interaural

[3] `intelligentheadset.com`

cues presented in Section 2.4.2 into the binaural output signals. Given a desired source in the far-field at a lateral angle, γ, and approximating the listener's ears by two points in free space, the propagation path difference, Δs, from the source to the two ears can be approximated by the *sine law* (Blauert, 1996):

$$\Delta s = d \sin \gamma, \tag{2.4}$$

where d is the distance between the two points. Using this simple approximation, the interaural time difference (ITD), τ_{itd} can be calculated as

$$\tau_{\text{itd}} = \frac{s}{c} = \frac{d \sin \gamma}{c}, \tag{2.5}$$

where c denotes the speed of sound. Therefore, to render a monophonic source at a lateral angle, γ, the ear input signal at the contralateral ear should be delayed by τ_{itd} with respect to the ipsilateral ear input signal. Non-negative delays, $\tau(\gamma)$, that can be applied to each ear input signal to yield an ITD approximately equal to τ_{itd} can be calculated as follows (Pulkki et al., 2011):

$$\tau(\gamma_{\text{c}}) = \begin{cases} \frac{a}{c} \cdot (1 - \cos(\gamma_{\text{c}} + \frac{\pi}{2})) & \text{if } |\gamma_{\text{c}} + \frac{\pi}{2}| < \frac{\pi}{2}, \\ \frac{a}{c} \cdot (|\gamma_{\text{c}} + \frac{\pi}{2}| - \frac{\pi}{2} + 1) & \text{else,} \end{cases} \tag{2.6}$$

where a denotes the effective head radius (Pulkki et al., 2011), and γ_{c} is the channel-dependent lateral angle in radians: $\gamma_{\text{c}} = \gamma$ for the left ear, and $\gamma_{\text{c}} = -\gamma$ for the right ear. This simple ITD approximation has proven effective in practical applications, though more sophisticated models have been proposed in the literature (Duda et al., 1999; Minnaar et al., 2000). The interaural level difference (ILD) of a source as a function of the lateral angle, γ, can be approximated by a simple infinite impulse response (IIR) filter (Pulkki et al., 2011):

$$H_{\text{hs}}(z, \gamma_{\text{c}}) = \frac{(\frac{c}{a} + \alpha(\gamma_{\text{c}}) f_{\text{s}}) + (\frac{c}{a} - \alpha(\gamma_{\text{c}}) f_{\text{s}})\, z^{-1}}{(\frac{c}{a} + f_{\text{s}}) + (\frac{c}{a} - f_{\text{s}})\, z^{-1}}, \tag{2.7}$$

with

$$\alpha(\gamma_{\text{c}}) = 1.05 + 0.95 \cos\left(\frac{180}{150}\left(\gamma_{\text{c}} + \frac{\pi}{2}\right)\right), \tag{2.8}$$

where f_{s} denotes the audio sampling rate.

2.5.3 Rendering polar angle: Spectral cues

To render the polar angle (or elevation) of a sound source, appropriate spectral cues have to be encoded in the ear input signals, as discussed in Section 2.4.3. Algazi et al. (2002) propose the use of simple geometric models of the torso and head to obtain polar-angle dependent acoustic cues at low frequencies. Other approaches to model the effect of head, torso, and pinnae on the sound

reaching the ears include the use electroacoustic filters tuneable according to a set of anthropometric measures of the listener (Genuit, 1987), and numerical approximations of the HRTF using finite difference (Xiao and Huo Liu, 2003) and boundary element methods (Katz, 2001; Gumerov et al., 2010).

2.5.4 Rendering distance and reverberation

Manipulating sound intensity provides a straightforward cue for source distance (see Section 2.4.4). Sound travelling in air is attenuated due to air absorption, with high frequencies attenuated the most (Zahorik et al., 2005). For broadband signals, adjusting the relative sound intensity at high frequencies can provide a distance cue (Zahorik et al., 2005). The ILD boost of sources in the nearfield (see Section 2.4.4) can be approximated via a range-dependent spherical head model (Duda and Martens, 1998; Spagnol et al., 2012). For sources in reverberant virtual environments, adjusting the direct-to-reverberant ratio according to the source distance provides a crucial cue for distance perception (Zahorik et al., 2005). Bronkhorst and Houtgast (1999) introduced a model relating the perceived source distance to the ratio between direct and reverberant energy. The model was later updated to explain the effect of lateral room reflections on the perceived distance (Bronkhorst, 2002). Rendering room reflections via artificial reverberation (Välimäki et al., 2012) and encoding interaural and spectral cues in each reflection yields a simulated binaural room impulse response (BRIR). The BRIR captures the effect of both the room and the listener on the sound. Using a simulated BRIR to add reverberation to a virtual source allows to affect the perceived source distance by adjusting the direct-to-reverberant ratio (Bronkhorst and Houtgast, 1999; Kolarik et al., 2013), the number of lateral reflections (Bronkhorst, 2002), and the temporal envelope of the BRIR (Albrecht and Lokki, 2013).

Kan et al. (2011) proposed a method for synthesising BRIRs from B-format recordings and HRTF measurements. Gamper and Lokki (2011) proposed a method for obtaining in-situ BRIRs from the microphone signals of a binaural AAR headset (see Fig. 2.2). When the listener snaps a finger, the response is recorded at the headset microphones. The recording of the impulse-like finger snap directly yields a coloured estimate of the in-situ BRIR. A block diagramme of the proposed approach is shown in Fig. 2.3.

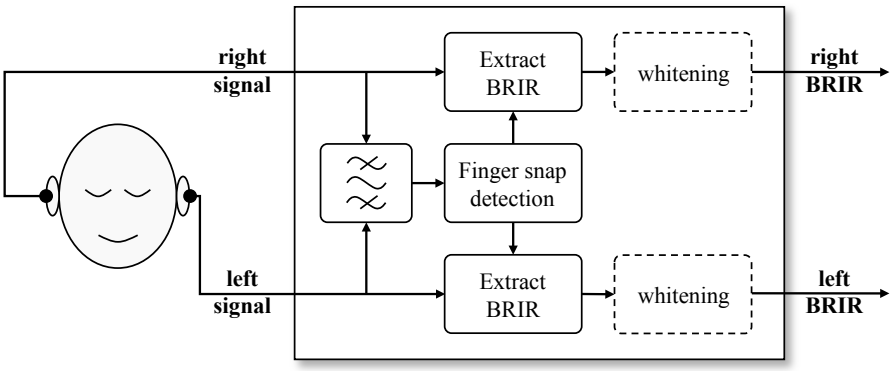

Figure 2.3. BRIR extraction from binaural AAR headset microphones (Gamper and Lokki, 2011)

2.5.5 Rendering using head-related transfer functions

A straight-forward way to encode interaural and spectral cues of a virtual source into the ear input signals is to filter the signals with a pair of head-related transfer functions (HRTFs) corresponding to the desired source direction. The filtering can be performed via convolution in the time-domain (Zotkin et al., 2004), or as a complex multiplication in the frequency-domain (Smith, 2007). While the frequency-domain approach may reduce the computational complexity of the filtering (Smith, 2007), it has an inherent input-to-output delay: The output of the frequency-domain filtering is only available after processing the whole input signal. In contrast, time-domain filtering produces a valid output sample for every new input sample (Zotkin et al., 2004). To reduce the delay of frequency-domain filtering it is typically performed on blocks of the input signal (Zotkin et al., 2004). The output signal can then be obtained by combining the output blocks of the frequency-domain filter using an *overlap-add* or *overlap-save* scheme (Smith, 2007).

Filtering a sound signal with an appropriate set of HRTFs yields ear input signals for rendering a virtual source at the direction defined by the HRTFs. To render a virtual source with high fidelity, the ear input signals should closely match the ear signals produced by a real source. This requires that the HRTFs used for filtering closely match the listener's own HRTFs.

Measuring HRTFs on a human test subject is a complex and time-consuming process. The measurement is typically performed in an anechoic chamber, by recording the ear input signals of a sound emitted from various locations around the listener. A large number of measurement locations is necessary to record HRTFs with sufficient spatial resolution. Prior studies suggest

measurements be taken at elevation intervals of 5–15 degrees, with 4–5 degrees azimuthal spacing on the horizontal plane and sparser measurements towards extreme elevations (Zhong and Xie, 2009; Zhang et al., 2012). To capture near-field HRTFs, these measurements would have to be performed at various distances (Brungart, 2002), resulting in thousands of measurement locations. Therefore, instead of using measured individual HRTFs, practical applications often rely on generic HRTF sets (Gardner and Martin, 1995; Algazi et al., 2001a; IRCAM, 2013). However, the use of nonindividual HRTFs, whether from another human subject or from a dummy head, can deteriorate the localisation performance of the listener (Wenzel et al., 1993; Møller et al., 1996; Møller et al., 1999). A study by Jin et al. (2000) indicates that accurate localisation requires about 60 percent of individual differences between test subjects to be preserved. Approaches have been proposed to select suitable HRTFs from a measurement set based on the listener's preference (Katz and Parseihian, 2012) or anthropometric features (Jin et al., 2000; Zotkin et al., 2003; Schönstein and Katz, 2010; Katz and Schönstein, 2013), and to numerically approximate individual HRTFs based on a geometric model of the listener (Katz, 2001; Xiao and Huo Liu, 2003; Gumerov et al., 2010). Experiments by Parseihian and Katz (2012) indicate that listeners may adapt to nonindividual HRTFs after a training period.

If measured HRTFs are used in spatial rendering, they are usually available only for certain directions. HRTF measurements are typically performed at a fixed distance from the test subject on a discrete measurement grid. To render a virtual source at a direction not available in the measurement set, a suitable pair of HRTFs for the desired direction has to be estimated from the available measurements. This can be done via *HRTF interpolation*, a technique that is discussed in Chapter 5.

When using headphones for playback, equalisation should be applied to flatten the frequency response of the playback system and thus minimise its effect on the binaural signals (Zahorik et al., 1995). Kim and Choi (2005) argue for the use of individual equalisation filters, to account for individual differences between listeners.

2.5.6 Rendering dynamic cues

To support interaction of the listener with a virtual auditory environment, the virtual sound sources should respond to listener movement in a similar way as real sound sources would. This requires both measuring the listener's position and orientation (via *motion tracking*, see Chapter 3) and encoding

dynamic cues into the ear input signals. Dynamic cues arise implicitly from a change of the localisation cues encoded into the ear input signals when updating the position of a virtual sound source in accordance with a change in the position and/or orientation of the listener. Dynamic cues can improve the localisation and perceived quality of spatialised audio. To render dynamic cues accurately, the rendering system should have a system delay smaller than 500 ms (Wenzel, 1999; Wenzel, 2001; Yairi et al., 2008) and an update rate higher than 18 Hz (Laitinen et al., 2012).

2.5.7 Properties and limitations of spatial rendering

The goal of rendering spatial sound for augmented reality is to embed virtual sounds into the natural acoustic environment. This implies that the rendering system should allow the precise placement of a virtual source. A real sound source usually causes the perception of an auditory event that lies at or close to the source position (Blauert, 1996). However, the same may not be true for a virtual sound source. A common problem of spatialised audio is inside-the-head locatedness (IHL) (Blauert, 1996). IHL occurs when a virtual sound source is perceived as emanating from inside the head, i.e., the auditory event caused by the virtual source resides somewhere between the ear entrances. Related to IHL is the concept of *externalisation* (Kim and Choi, 2005), that describes how well a listener perceives a virtual sound source to emanate from outside the head. Ideally, a perfectly externalised source would be indistinguishable from a real source (Hartmann and Wittenberg, 1996). However, rendering an externalised source via headphones is a challenging problem. In previous studies, rendering a virtual source that is indistinguishable from a real one has been achieved with careful calibration of the rendering system. Probe microphones were inserted into the ear canals of a test subject to measure the ear input signals when exposed to a real source. Using the recorded ear input signals as a baseline, the study authors were able to render a virtual source that the test subject would confuse with a real one (Zahorik et al., 1995; Hartmann and Wittenberg, 1996; Langendijk and Bronkhorst, 2000; Härmä et al., 2004). However, the illusion of the virtual source being a real one could only be created with certain sound samples (Härmä et al., 2004), and it vanished if the rendering introduced errors in the phase or ILD of the ear input signals (Hartmann and Wittenberg, 1996). Experiments by Begault et al. (2000) indicate that reverberation increases the perceived externalisation of a virtual source. Kim and Choi (2005) state that the use of individual HRTFs and headphone equalisation improves the perceived externalisation of virtual sources in the horizontal plane, except for

sources straight ahead.

Rendering a well-externalised virtual source in front of the listener is difficult, especially if no visual cues are present that correlate with the auditory event (Wenzel et al., 1993). Therefore, virtual sources straight ahead are particularly prone to IHL as well as *front–back confusions*, where a source in front is perceived to be positioned behind the listener (Wenzel et al., 1993). Front–back confusions occur due to the ambiguities of interaural cues and the resulting cone of confusion (see Section 2.4.3) (Wenzel et al., 1993). To lower front–back confusion rates, the use of individual HRTFs has been suggested (Wenzel et al., 1993). Furthermore, dynamic cues induced by head movements allow listeners to determine whether a source is in the front or in the back (Wenzel et al., 1993; Begault et al., 2000).

3. Motion tracking

In the context of human–computer interaction in general, and AR in particular, knowing the position and orientation of the user allows to enhance how the user interacts with and perceives the real environment. The country the user is located in can serve as an indicator for the language in which information or user-interface elements should be presented. The approximate geographic location can be used to tailor the displayed information for the specific environment the user is in, for instance to point out nearby friends (Yu et al., 2011). Combining information about the geographic location with head orientation data allows overlaying information onto the physical environment (Feiner et al., 1997). Precise position and orientation data at a high update rate enables the creation and control of immersive and interactive augmented environments (Zimmermann and Lorenz, 2008). Given that the requirements regarding the availability and the temporal and spatial resolution of position and orientation data vary between applications, a variety of motion tracking methods and systems have been developed to serve those requirements (Hightower and Borriello, 2001; Welch and Foxlin, 2002).

In Publications I and II, methods are proposed for tracking the head orientation and position of human speakers in a collaborative AR environment, such as the one presented by Butz et al. (1999), or a teleconference. The approaches take advantage of binaural AAR headsets worn by the users, as depicted in Fig. 2.2. The headsets function both as the playback system for delivering AAR content and as sensors for the proposed acoustic tracking system. No other sensors or markers need to be worn by the users. Knowing the position and orientation of users in a collaborative environment allows embedding virtual auditory objects or information into the shared space as an AAR overlay. An example for the use of AAR content in a collaborative environment or conference is the ability to render a remote participant acoustically at a stable position in the shared environment. Due to the “cocktail party effect” (see

Section 2.2), spatial rendering of a teleconference participant would potentially enhance the intelligibility, especially if there are several remote participants. Next, an overview of existing tracking technologies is presented.

3.1 Tracking techniques and systems

AR systems for outdoor applications often include a GPS sensor to retrieve the geographic location of a user (Feiner et al., 1997; Julier et al., 2000; Azuma et al., 2001; Härmä et al., 2004; Reitmayr and Drummond, 2006). The advantage of using a GPS sensor for tracking is that it is portable, it provides absolute location information, and it is often readily available in portable devices, e.g., smart-phones. On the downside, GPS performance depends on the signal strength of GPS satellites (Reitmayr and Drummond, 2006), which is why it can usually not be used indoors (Zeimpekis et al., 2002). Other approaches to determine the geographic location of a user outdoors include mobile network cell identification (Zeimpekis et al., 2002), Wireless Local Area Network (WLAN) positioning (Anisetti et al., 2011), and methods based on measuring the cellular-network signal attenuation (Anisetti et al., 2011).

For outdoor AR applications that require tracking the azimuth angle of the user's head, a digital compass can be used (Glumm et al., 1998; Hoff and Azuma, 2000). While a digital compass has the advantage of providing absolute orientation data, it typically suffers from slow update rate and high latency (Azuma et al., 1999). Therefore, outdoor AR systems often combine a compass with *inertial sensors* (Welch and Foxlin, 2002), i.e., accelerometers and gyroscopes (Azuma et al., 1999; Reitmayr and Drummond, 2006). The gyroscopes track rotation in 3-D. Double integration of the accelerometer data yields a position estimate (Welch and Foxlin, 2002), and the constant acceleration due to gravity can be used to determine the orientation relative to the gravity vector. Inertial sensors provide the advantage that they are self-contained and thus require no external infrastructure, such as satellites or network base stations (IEEE, 2001; Welch and Foxlin, 2002). However, the absence of an external reference makes inertial sensors prone to drift (DiVerdi and Höllerer, 2007).

Modern smartphones typically come equipped with inertial sensors, a compass, and a GPS receiver (Li et al., 2013). When using a smartphone to deliver AR content, the drift of the fast, high-resolution inertial tracking can be corrected using the coarse, drift-free GPS and compass tracking (DiVerdi and Höllerer, 2007). Furthermore, the camera of an AR system can be used to estimate

device motion from the camera stream's optical flow (DiVerdi and Höllerer, 2007; Li et al., 2013), or from natural features of the environment (Schmalstieg et al., 2011), thus further improving tracking accuracy.

Outdoor AR applications typically require self-contained tracking methods that do not rely on external references (Azuma et al., 1999). Indoor applications, on the other hand, may take advantage of an environment that is prepared for the specific application (Azuma et al., 1999). Motion tracking in prepared environments may employ external hardware to track the user, or as reference for a mobile tracking system. As an example, the "Active Badge" tracking system (Want et al., 1992) uses infrared detectors mounted on walls and ceilings of a large office building to determine the location of badges emitting infrared pulses. Another example of tracking in prepared environments is fiducial tracking, whereby the position and orientation of a camera is estimated relative to known markers (Kato and Billinghurst, 1999; Welch and Foxlin, 2002). The markers can either be placed on the user or device to be tracked, i.e., "outside-looking-in" (Welch and Foxlin, 2002), or distributed in the environment to track the motion of a user-worn camera, i.e., "inside-looking-out" (Welch and Foxlin, 2002). Examples of fiducial markers are AR markers of known shape and size (Kato and Billinghurst, 1999), reflective markers (Chung et al., 2001), or light-emitting diodes (Foursa, 2004). While optical tracking systems can be quite complex and expensive, a major advantage of fiducial tracking using AR markers is the minimal hardware cost. Position and orientation tracking in 3-D can be implemented using markers printed on paper and a single camera (Kato and Billinghurst, 1999), allowing to run AR applications on a basic cellphone with integrated camera (Möhring et al., 2004).

Magnetic tracking systems measure the field produced by magnetic coils to estimate the position and orientation of a magnetic sensor (Welch and Foxlin, 2002). Unlike camera-based tracking systems, magnetic tracking does not require a line of sight (Welch and Foxlin, 2002). Due to their high accuracy but limited range, magnetic trackers are often used to track the head of a user for spatial sound rendering (Wenzel, 1999; Macpherson and Middlebrooks, 2002; Parseihian and Katz, 2012).

Acoustic tracking systems estimate position and orientation by analysing sound waves. As with camera-based systems, both "outside-in" and "inside-out" approaches exist, and the wavelengths used may be within or outside the human-perceptible range. Tracking can be performed by estimating the position of either a sound receiver or a sound emitter (Hightower and Borriello, 2001). If active sound emitters, including loudspeakers, are used, the tracking

method	position tracking	orientation tracking	self-contained	accuracy	update rate
GPS	absolute	-	no	low	low
base station	absolute	-	no	low	low
WLAN	absolute	-	no	low	low
compass	-	absolute	yes	medium	medium
accelerometer	relative	absolute	yes	medium	high
gyroscope	-	relative	yes	high	high
optical flow	-	relative	yes	high	medium
active badge	absolute	-	no	low	low
fiducial	absolute	absolute	no	high	medium
magnetic	absolute	absolute	no	high	high
acoustic	absolute	absolute	no	high	high

Table 3.1. Overview of tracking methods, adapted from Hightower and Borriello (2001); DiVerdi and Höllerer (2007); Li et al. (2013).

system typically operates in the ultrasonic frequency range (Welch and Foxlin, 2002). The emitters produce short sound pulses that are used to estimate the times of arrival (TOAs) at the receivers (Ward et al., 1997). Using 3-D ultrasound imaging, medical instruments can be tracked with high precision by attaching a passive marker to them (Novotny et al., 2007). Acoustic tracking in the audible frequency range can be used to track a human speaker's head orientation (Tikander et al., 2004; Lacouture-Parodi and Habets, 2012, 2013) or location (Ward et al., 2003; Tikander et al., 2004; Pertilä et al., 2008; Wu et al., 2013; Schwartz and Gannot, 2014; Zhong et al., 2014). An advantage of acoustic tracking over other tracking methods is that the user's position can be tracked without her or him wearing any sensors or markers. Table 3.1 presents an overview of tracking methods and their respective properties.

3.2 Acoustic tracking with particle filtering

The tracking approaches proposed in Publications I and II rely on *time-delay estimation* (Knapp and Carter, 1976; Ward et al., 2003), whereby the speaker location and orientation is inferred from time of arrival (TOA) and time-difference of arrival (TDOA) estimates. Next, the general framework for speaker location tracking is introduced. In Section 3.4, the adaptation of this

framework for speaker orientation tracking is presented.

The TOA, t_i, of a receiver at the location, $\mathbf{r}_i$, is the delay a sound signal incurs when travelling from the source position, $\mathbf{s}$, to the receiver:

$$t_i = c^{-1}\|\mathbf{s} - \mathbf{r}_i\|, \tag{3.1}$$

where c denotes the speed of sound, and $\|\cdot\|$ is the Euclidean norm. The TDOA, $\tau_{i,j}$ between receivers i and j, is the difference between their respective TOAs:

$$\tau_{i,j} = t_i - t_j = c^{-1}\|\mathbf{s} - \mathbf{r}_i\| - c^{-1}\|\mathbf{s} - \mathbf{r}_j\|. \tag{3.2}$$

A TDOA estimate, $\hat{\tau}_{i,j}$, can be obtained via the generalised correlation framework with phase transform (Knapp and Carter, 1976):

$$R_{i,j}(\tau) = \int \frac{X_i(f)X_j^*(f)}{|X_i(f)X_j^*(f)|} e^{j2\pi f\tau} df, \tag{3.3}$$

where $X_i(f)$ is the Fourier transform of the microphone signal, $x_i(t)$, recorded at the ith receiver, and $(\cdot)^*$ denotes the complex conjugate. With Eq. (3.3), the TDOA estimate, $\hat{\tau}_{i,j}$, is obtained as

$$\hat{\tau}_{i,j} = \arg\max_{\tau} \left(R_{i,j}(\tau)\right). \tag{3.4}$$

Note that the TDOAs can be estimated from the microphone signals without knowledge of the source or receiver positions. With known receiver positions, $\mathbf{r}_i, \mathbf{r}_j$, the estimated TDOA, $\hat{\tau}_{i,j}$, yields a locus of potential source locations. In absence of signal reflections and noise, the loci of different receiver pairs intersect at the true source location, $\mathbf{s}$. However, this is usually not true in the presence of reflections or noise (Ward et al., 2003). Therefore, for practical applications, a different approach is needed. Here, the likelihood of a set of candidate source positions is calculated, and a source position estimate is derived from this set rather than directly from the TDOA estimates.

3.2.1 Likelihood function

For receivers at locations, $\mathbf{r}_i$, $\mathbf{r}_j$, and a candidate source position, $\mathbf{s}$, the *expected* TDOA, $\tau_{i,j}$, and the *estimated* TDOA, $\hat{\tau}_{i,j}$, are given via Eqs. (3.2) and (3.4), respectively. The likelihood of observing $\hat{\tau}_{i,j}$ for the source position, $\mathbf{s}$, can be expressed as (Lehmann, 2004)

$$p(\tau_{i,j}(\mathbf{s})|\hat{\tau}_{i,j}, \sigma_{i,j}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\tau_{i,j}(\mathbf{s}) - \hat{\tau}_{i,j})^2}{2\sigma_{i,j}^2}\right), \tag{3.5}$$

i.e., a normal distribution with variance, $\sigma_{i,j}^2$, and mean, $\hat{\tau}_{i,j}$. The variance is assumed to be equal for all microphone pairs, and the estimation errors for

each pair are assumed to be statistically independent. The total likelihood of observing all estimated TDOAs for a candidate source position, $\mathbf{s}$, can be calculated via the *maximum likelihood estimation (MLE)* function (Lehmann, 2004)

$$p(\mathbf{s}) = \prod_{\{i,j\}=1}^{\mathrm{M}} p(\tau_{i,j}(\mathbf{s})|\hat{\tau}_{i,j}, \sigma_{i,j}), \tag{3.6}$$

where M denotes the number of receiver pairs. By evaluating Eq. (3.6) for a set of candidate source positions, a source position estimate, $\hat{\mathbf{s}}$, can be derived via *particle filtering* (Ward et al., 2003; Lehmann, 2004), a technique that is briefly introduced in the next section.

3.2.2 Particle filtering

Given an array of microphones, acoustic tracking can be described in terms of a Bayesian filtering problem, where an estimate for the current source position (and velocity) is obtained from a posterior probability density function (PDF) based on all localisation information available for the source up to the current time step (Lehmann et al., 2007). A particle filter provides a solution to the Bayesian filtering problem by approximating the posterior PDF by a set of particles and associated weights (Lehmann et al., 2007). The particle locations constitute the candidate source positions to be evaluated. To initialise the particle filter, a set of K particles is uniformly distributed in the tracking area. The filtering is implemented in three steps: *prediction*, *update*, and *resampling*. In the prediction step, the particle locations are propagated according to a model of the source dynamics (Ward et al., 2003). Here, Brownian motion is assumed as the dynamic model, that is, the particles are propagated according to a random distribution (Pertilä et al., 2008). In the update step, a weight, w_k, is calculated for each particle at location, $\mathbf{p}_k$, with Eq. (3.6) as

$$w_k = p(\mathbf{p}_k). \tag{3.7}$$

The weights, w_k, are normalised so that

$$\sum_{k}^{\mathrm{K}} \widetilde{w}_k = 1, \tag{3.8}$$

where

$$\widetilde{w}_k = w_k \left(\sum_{k}^{\mathrm{K}} w_k \right)^{-1}. \tag{3.9}$$

The source location estimate, $\hat{\mathbf{s}}$, is given as the weighted sum of the particle locations, $\mathbf{p}$:

$$\hat{\mathbf{s}} = \sum_{k}^{\mathrm{K}} \widetilde{w}_k \mathbf{p}_k. \tag{3.10}$$

In the resampling step, a fixed number of particles are redrawn from the particle set according to their weights (Lehmann, 2004). Particles with low weights are discarded and replaced by particles with higher weights. In Publications I and II, *stratified resampling* is used (Douc and Cappé, 2005).

3.3 Tracking speaker position

Acoustic source tracking setups typically consist of several microphones, distributed across the room (Ward et al., 2003; Pertilä et al., 2008; Cho et al., 2010) or arranged in clusters or arrays (Sun et al., 2009; Talantzis, 2010). This allows reliable tracking of the speakers, given that the acoustic conditions are favourable (Pertilä et al., 2008). Tracking the position and head orientation of a user via binaural headset microphones was previously proposed using anchor sound sources at known positions (Tikander et al., 2004). For reliable speaker tracking, the aforementioned systems require the installation of multiple arrays, which can be complex and costly, or anchor sources at known positions.

The position tracking system proposed in Publication II relies on a single microphone array and a binaural AAR headset (such as the one depicted in Fig. 2.2) worn by the users. The advantage of integrating user-worn microphones into the tracking system is their vicinity to the acoustic source, i.e., the speaker, which in turn can result in better signal-to-noise ratio (SNR) and hence better raw data for the acoustic source tracking. Furthermore, assuming the distance between the user-worn microphones and the speaker to be constant, the speaker–array and speaker–listener distances can be estimated. The tracking system proposed in Publication II takes advantage of distance estimates obtained from the headset microphones for improved tracking accuracy and robustness. An overview of the tracking setup is depicted in Fig. 3.1.

The proposed approach consists of three parts. First, basic voice activity detection is performed to determine the active speaker from the binaural microphone signals. Then, the position of the active speaker is tracked via particle filtering, using the framework described in Section 3.2. Finally, the distance of each conferee to the active speaker is estimated to derive an importance function for prior weighting of the particles of silent conferees.

3.3.1 Voice activity detection

For the purpose of determining who spoke and when, a basic voice activity detection is implemented. It relies on thresholding the signal energy recorded

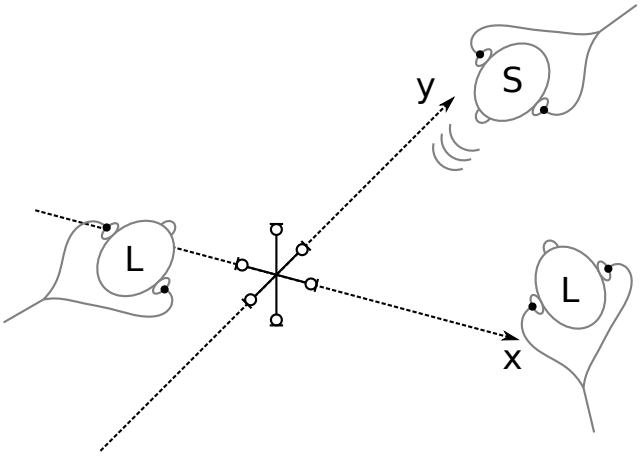

Figure 3.1. Position tracking setup. The speaker and listeners are denoted by S and L, respectively.

at the binaural headset microphones and the tracking evidence found for the particles of each conferee. This simple approach worked well for the given setup, though more sophisticated voice activity detection methods exist (Sohn et al., 1999).

3.3.2 Time-delay estimation and likelihood function

Applying the definitions used in Section 3.2 to the speaker position tracking scenario, the source at position, $\mathbf{s}$, corresponds to the human speaker, and the receivers at positions, $\mathbf{r}_i$, consist of the microphones in the reference microphone array and the binaural headsets. TDOA estimates between pairs of reference-array and binaural-headset microphones are obtained via Eq. (3.4). The expected TDOA, τ, between a microphone of the reference microphone array at position, $\mathbf{r}_i$, and the speaker's left and right binaural headset microphones at locations, $\mathbf{r}_{\mathrm{spL}}$, $\mathbf{r}_{\mathrm{spR}}$, is given as

$$\begin{aligned} \tau_{i,\mathrm{spL}}(\mathbf{s}) &= c^{-1}\|\mathbf{s}-\mathbf{r}_i\| - c^{-1}\|\mathbf{s}-\mathbf{r}_{\mathrm{spL}}\|, \\ \tau_{i,\mathrm{spR}}(\mathbf{s}) &= c^{-1}\|\mathbf{s}-\mathbf{r}_i\| - c^{-1}\|\mathbf{s}-\mathbf{r}_{\mathrm{spR}}\|. \end{aligned} \tag{3.11}$$

Let d_{spE} denote the distance between the speaker's acoustic centre and left or right ear. It is assumed that d_{spE} is fixed and equal for both ears:

$$d_{\mathrm{spE}} = \|\mathbf{s}-\mathbf{r}_{\mathrm{spL}}\| = \|\mathbf{s}-\mathbf{r}_{\mathrm{spR}}\| \approx 0.18 \text{ m}. \tag{3.12}$$

Through substitution using Eqs. (3.1) and (3.12), the expected TDOA in Eq. (3.11) between a receiver in the reference microphone array at location, $\mathbf{r}_i$, and one of the speaker's binaural headset microphones at location, $\mathbf{r}_{\mathrm{spE}}$, becomes

$$\tau_{i,\mathrm{spE}}(\mathbf{s}) = t_i(\mathbf{s}) - c^{-1}d_{\mathrm{spE}}, \tag{3.13}$$

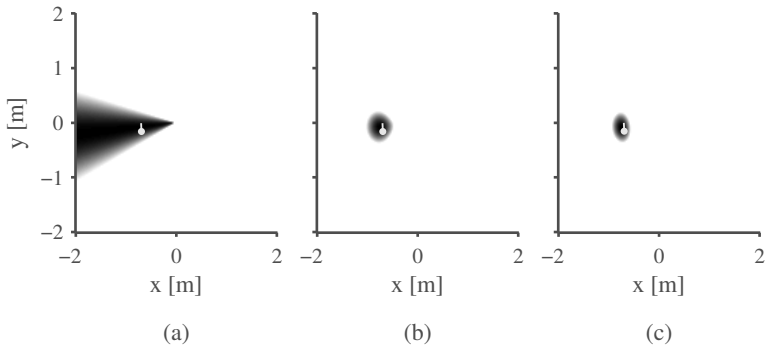

Figure 3.2. MLE function for tracking area, obtained from a reference microphone array at $(0, 0)$ and (a) no, (b) one, and (c) two binaural headset microphones worn by the speaker. The light-grey dot indicates the true position and head orientation of the speaker.

By applying the relation between TOA and TDOA given in Eq. (3.13), a TOA estimate, $\hat{t}$, for a receiver in the microphone array at location, $\mathbf{r}_i$, can be derived from a TDOA estimate, $\hat{\tau}$, between the receiver and the speaker's jth binaural headset microphone as

$$\hat{t}_{i,\mathrm{sp}j} = c^{-1} d_{\mathrm{spE}} + \hat{\tau}_{i,\mathrm{sp}j}, \tag{3.14}$$

where $j = 1$ denotes the speaker's left and $j = 2$ the speaker's right binaural microphone. With Eq. (3.14), the distance from the ith receiver to the speaker can be estimated from the TDOA estimate between the receiver and the speaker's jth binaural microphone as

$$\hat{d}_{i,\mathrm{sp}j} = c\,\hat{t}_{i,\mathrm{sp}j} = d_{\mathrm{spE}} + c\,\hat{\tau}_{i,\mathrm{sp}j}. \tag{3.15}$$

With expected and estimated TOAs given by Eqs. (3.1) and (3.14), respectively, the MLE function in Eq. (3.6) can be expanded to take advantage of the speaker distance estimates, $\hat{d}$. The MLE function for a combination of binaural headset microphones worn by the speaker, N reference array microphones, and a total of M microphone pairs, is given as

$$p(\mathbf{s}) = \underbrace{\left(\prod_{\{i,j\}=1}^{\mathrm{M}} p(\tau_{i,j}(\mathbf{s}) | \hat{\tau}_{i,j}, \sigma_{i,j}) \right)}_{\text{MLE using TDOA estimates}} \times \underbrace{\left(\prod_{i=1}^{\mathrm{N}} \prod_{j=1}^{2} p(t_i(\mathbf{s}) | \hat{t}_{i,\mathrm{sp}j}, \sigma_{i,\mathrm{sp}j}) \right)}_{\text{MLE using TOA estimates}}. \tag{3.16}$$

An example of the MLE function computed over the tracking area with no, one, and two binaural headset microphones is shown in Figure 3.2.

3.3.3 Listener importance function

The listener importance function is used to calculate particle weights for the listeners, i.e., the users that are silent while the speaker is talking. The listener

particle weights indicate where a listener's particles are to be sampled once she or he starts talking and tracking resumes (Lehmann, 2004).

In analogy to Eq. (3.13), the expected TDOA, τ, between the listener's jth binaural headset microphone and one of the speaker's headset microphones is given as

$$\tau_{\mathrm{lis}j,\mathrm{spE}}(\mathbf{s}) = t_{\mathrm{lis}j}(\mathbf{s}) - c^{-1} d_{\mathrm{spE}}. \tag{3.17}$$

The estimated TOA, $\hat{t}$, for the speaker's ith and the listener's jth binaural headset microphone is given in analogy to Eq. (3.14) as

$$\hat{t}_{\mathrm{sp}i,\mathrm{lis}j} = c^{-1} d_{\mathrm{spE}} + \hat{\tau}_{\mathrm{sp}i,\mathrm{lis}j}, \tag{3.18}$$

where $\hat{\tau}_{\mathrm{sp}i,\mathrm{lis}j}$ is the estimated TDOA obtained via Eq. (3.4). The distance between the speaker and the listener's jth binaural microphone can be estimated as

$$\hat{d}_{\mathrm{sp}i,\mathrm{lis}j} = c\,\hat{t}_{\mathrm{sp}i,\mathrm{lis}j} = d_{\mathrm{spE}} + c\,\hat{\tau}_{\mathrm{sp}i,\mathrm{lis}j}. \tag{3.19}$$

Taking advantage of the speaker–listener distance estimate, the listener importance function is given as an MLE function via

$$p_{\mathrm{I}}(\mathbf{s}) = \underbrace{\left(\prod_{i=1}^{\mathrm{N}} \prod_{j=1}^{2} p(\tau_{i,\mathrm{lis}j}(\mathbf{s}) | \hat{\tau}_{i,\mathrm{lis}j}, \sigma_{i,\mathrm{lis}j}) \right)}_{\text{MLE using TDOA estimates}} \times \underbrace{\left(\prod_{i=1}^{2} \prod_{j=1}^{2} p(t_{\mathrm{sp}i,\mathrm{lis}j}(\mathbf{s}) | \hat{t}_{\mathrm{lis}j}, \sigma_{\mathrm{sp}i,\mathrm{lis}j}) \right)}_{\text{MLE using TOA estimates}}. \tag{3.20}$$

Figure 3.3 illustrates an example of the MLE function computed over the tracking area for one listener with both speaker and listener wearing no, one, and two binaural headset microphones. As shown in Fig. 3.3(a), no listener importance function can be derived without user-worn headset microphones. With at least one user-worn microphone per conferee, the importance function has the form of a circle, centred at the estimated speaker location, with a radius corresponding to the estimated speaker–listener distance. Using just one binaural headset microphone, the head orientation of the listener introduces a bias of max. ±0.1 m (i.e., half the head radius) to the estimated speaker–listener distance (see Fig. 3.3(b)).

3.3.4 Particle filtering

The particle filtering approach used for the speaker location tracking framework is introduced in Section 3.2. Each user is tracked by a separate particle filter. The particle filters are initialised by distributing the particles uniformly in the tracking area. In the update step, the particle weights for the active speaker are calculated via Eq. (3.7), using the MLE function given in Eq. (3.16). The

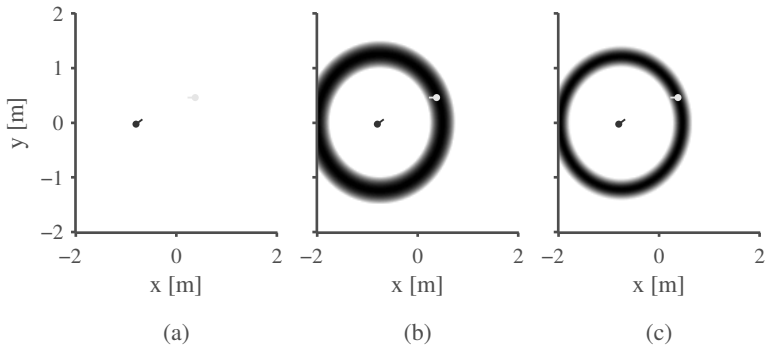

Figure 3.3. Importance function for listener P2 with (a) no, (b) one, and (c) two user-worn microphones. It is derived from the estimated distance to the active speaker P1 and used for prior weighting of the particles of P2. The dark-grey and the light-grey dot indicate the true position and head orientation of P1 and P2, respectively.

listener particles are updated using the listener importance function given in Eq. (3.20). In the resampling step, stratified resampling is applied to all particle filters. A speaker location estimate, $\hat{\mathbf{s}}$, is obtained as a weighted sum of the speaker's particle locations via Eq. (3.10).

3.4 Tracking listener orientation

Tracking the head orientation of users in an AAR environment is necessary to render virtual auditory content overlaid onto the environment. Existing head-tracking systems are either camera-based or require attaching a sensor or marker to the user. In Publication I, a head-tracking algorithm is proposed that relies on binaural microphone signals recorded via an AAR headset worn by the user (see Fig. 2.2). Unlike previously proposed methods for tracking head orientation via binaural microphone signals (Tikander et al., 2004; Lacouture-Parodi and Habets, 2012, 2013), the approach presented in Publication I performs tracking solely based on the users' speech signals recorded via binaural headset microphones, and does not require anchor sources at known positions. The proposed approach relies on particle filtering, as described in Section 3.2, and assumes that the positions of the users are known.

3.4.1 Voice activity detection

The voice activity detection method used here is the same as described in Section 3.3.1, except that particle weights are not taken into account to determine the active speaker.

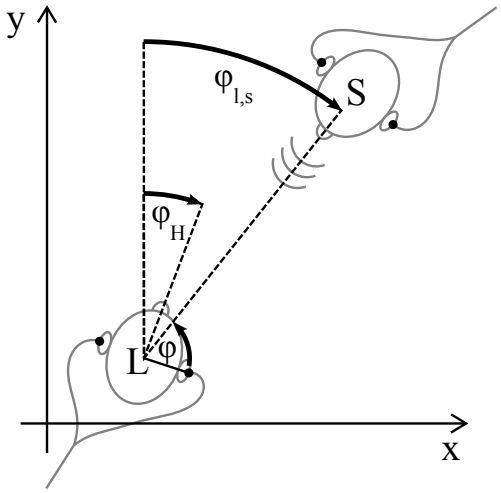

Figure 3.4. Schematic view of the head orientation estimation problem; the speaker and listener are denoted as S and L, respectively. The head orientation of the listener is denoted as φ_{H}.

3.4.2 Time-delay estimation and likelihood function

A schematic view of the orientation tracking problem is shown in Fig. 3.4. The proposed algorithm tracks the head orientation, φ_{H}, of the listener by estimating time delays between the speaker's and listener's binaural headset microphones.

Let φ denote the angle of incidence of the speech signal with respect to the listener's interaural axis, as depicted in Fig. 3.4, and $\mathbf{s}$ and $\mathbf{l}$ the speaker and listener positions, respectively. The speaker–listener distance, $\|\mathbf{s}-\mathbf{l}\|$, and the distance between the listener's jth binaural headset microphone and the speaker, $\|\mathbf{s}-\mathbf{r}_{\mathrm{lis}j}\|$, are related to φ by the law of cosines

$$\|\mathbf{s}-\mathbf{r}_{\mathrm{lis}j}\|^2 = a^2 + \|\mathbf{s}-\mathbf{l}\|^2 - 2a\|\mathbf{s}-\mathbf{l}\|\cos\varphi, \tag{3.21}$$

where a denotes the (listener's) head radius. With Eqs. (3.1) and (3.21), the expected TOA, t, for the listener's jth binaural headset microphone is

$$t_{\mathrm{lis}j}(\varphi) = c^{-1}\sqrt{a^2 + \|\mathbf{s}-\mathbf{l}\|^2 - 2a\cos\varphi\|\mathbf{s}-\mathbf{l}\|}. \tag{3.22}$$

Here, the speaker–listener distance, $\|\mathbf{s}-\mathbf{l}\|$, is assumed to be known. With Eq. (3.18), the estimated TOA, $\hat{t}$, for the speaker's ith and the listener's jth binaural headset microphone is obtained from TDOAs estimated using Eq. (3.4). The expected TDOA between the listener's two binaural headset microphones is given via Eq. (3.2) as

$$\tau_{\mathrm{lisL,lisR}}(\varphi) = t_{\mathrm{lisL}}(\varphi) - t_{\mathrm{lisR}}(\varphi). \tag{3.23}$$

The proposed orientation tracking method relies on calculating the likelihood of observing the estimated TOAs and TDOAs for a given angle of sound wave

incidence, φ, via a MLE function

$$p(\varphi) = p(\tau_{\text{lisL,lisR}}(\varphi)|\hat{\tau}_{\text{lisL,lisR}}, \sigma_{\text{lisL,lisR}}) \times \left(\prod_{i=1}^{2} \prod_{j=1}^{2} p(t_{\text{sp}i,\text{lis}j}(\varphi)|\hat{t}_{\text{lis}j}, \sigma_{\text{sp}i,\text{lis}j}) \right). \tag{3.24}$$

3.4.3 Particle filtering

The particle filtering for tracking the listener orientation is implemented analogously to the speaker location tracking framework presented in Section 3.3, with the difference that the particles track an angle rather than a position. The head orientations of all users are tracked by separate particle filters. The filters are initialised by distributing the particles uniformly between 0 and 2π. In the update step, the particle weights for the listeners, that is, all users except the currently active speaker, are calculated via Eq. (3.7):

$$w_k = p(\varphi_k), \tag{3.25}$$

where φ_k is the angle of the kth particle, and $p(\cdot)$ is the MLE function given in Eq. (3.24). An estimate for the speech signal's angle of incidence, $\hat{\varphi}$, with respect to the listener's interaural axis is given for each listener via Eq. (3.10) as

$$\hat{\varphi} = \sum_{k}^{\text{K}} \widetilde{w}_k \varphi_k, \tag{3.26}$$

where $\widetilde{w}$ is the normalised particle weight, as defined in Eq. (3.9). With Eq. (3.26), an estimate of the head orientation, $\hat{\varphi}_{\text{H}}$, with respect to the reference frame is obtained for each listener as

$$\hat{\varphi}_{\text{H}} = \varphi_{\text{l,s}} - \hat{\varphi} + \frac{\pi}{2}, \tag{3.27}$$

where $\varphi_{\text{l,s}}$ denotes the direction of the speaker relative to the listener (see Fig. 3.4). Here, $\varphi_{\text{l,s}}$ is derived assuming the positions of the listener and the speaker to be known. Alternatively, the position estimates obtained via the position tracking method presented in Publication II could be used to derive $\varphi_{\text{l,s}}$.

3.5 Experimental setup

In a case study, the positions and orientations of three conferees were tracked during 60 seconds of a conversation in a meeting scenario, as illustrated in Fig. 3.5. Tracking was implemented via binaural AAR headsets worn by each participant, and a reference microphone array located in the centre of the

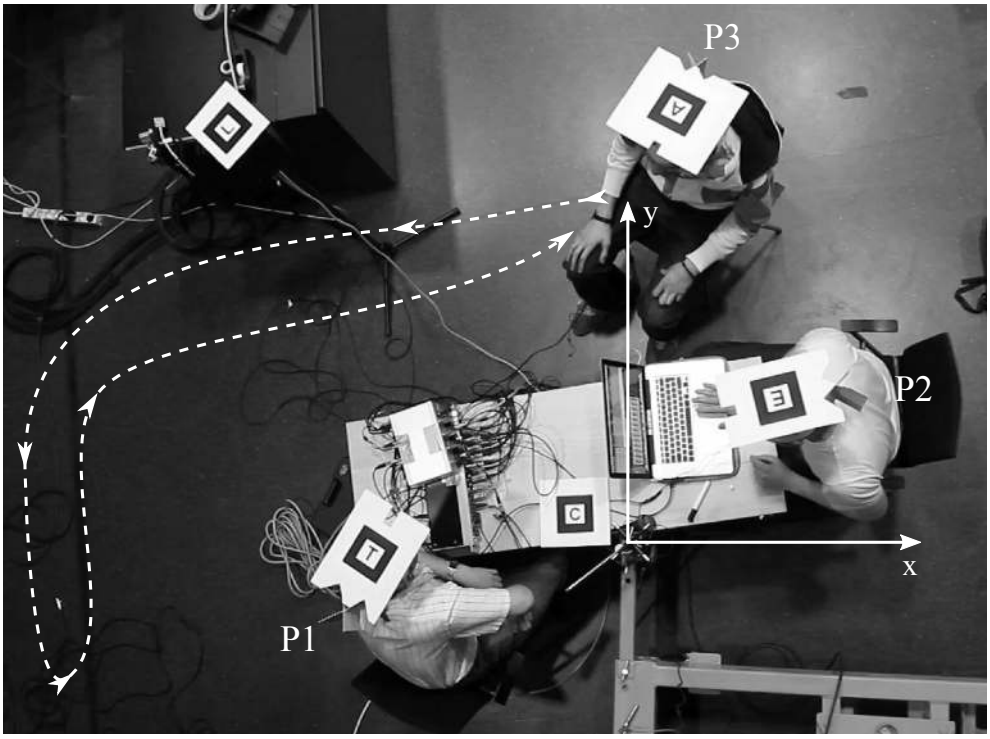

Figure 3.5. The experimental setup of the case study. The dashed line illustrates the path of conferee P3. The reference microphone array is located at the centre of the coordinate axes.

tracking area at $(0, 0)$. The experiment was conducted in a multipurpose space with a reverberation time of about 0.3 s (Kajastila et al., 2007) and an SNR between 15 and 30 dB. The ground truth data for the position and orientation tracking was obtained by tracking visually distinct markers placed on the head of each conferee using the ARToolkit, which for the given setup provides a location tracking accuracy of around 1 cm (Kato and Billinghurst, 1999) at an update rate of 30 Hz. Speech activity and the currently active speaker were determined in each frame using a simple voice activity detection method (see Sections 3.3.1 and 3.4.1). If speech activity was detected, the location of the active speaker and the orientations of the listeners were tracked. Speaker position and listener orientation tracking was implemented for each participant via particle filters with $\mathrm{K} = 100$ particles. The performance of the proposed head orientation tracking approach was compared to a reference method from the literature (Tikander et al., 2004).

	Distance RMSE [m]					Direction RMSE [deg]					Position RMSE [m]				
B	0	1	1*	2	2*	0	1	1*	2	2*	0	1	1*	2	2*
P1	5.53	0.07	0.11	**0.07**	0.08	8.0	10.7	9.5	8.5	**6.9**	5.54	0.15	0.15	0.13	**0.11**
P2	10.81	0.07	0.06	**0.05**	0.05	26.2	23.2	12.9	22.2	**11.5**	10.87	0.22	0.13	0.21	**0.12**
P3	1.61	**0.06**	0.09	0.09	0.07	10.4	15.7	**3.7**	7.1	3.7	1.64	0.39	0.12	0.14	**0.11**
mean	5.99	**0.07**	0.09	0.07	0.07	14.9	16.6	8.7	12.6	**7.4**	6.02	0.26	0.13	0.16	**0.11**

*Prior weighting of listener particles based on listener importance function.

Table 3.2. Speaker location tracking performance for the estimated distance, direction, and position of each conferee relative to the reference microphone array. B indicates the test condition, i.e., the number of binaural headset microphones used per conferee. The lowest RMSE for each conditions is shown in bold typeface.

Method	(Ward et al., 2003)	(Fallon and Godsill, 2010)	(Talantzis, 2010)	**here**
RMSE [m]	0.14	0.29	0.14	**0.11**

Table 3.3. Tracking performance compared to state of the art tracking systems tested under similar experimental conditions.

3.6 Results

3.6.1 Speaker location tracking

The root-mean square error (RMSE) for the speaker location tracking under various conditions is summarised in Table 3.2. When using the reference microphone array alone, tracking performance is poor due to the small microphone spacing of the array (see Table 3.2, B = 0). While the speaker direction estimation accuracy with the reference microphone array alone is comparable to the combination of array and binaural-headset microphones without listener importance functions, the distance estimation is substantially worse (see Table 3.2, B $\in \{1, 2\}$).

With the use of binaural headset microphones, the distance RMSE is below 0.09 m on average for all conditions, i.e., in a similar range as the head radii of the conferees (see Table 3.2, B > 0). This greatly improves the position tracking performance compared to using the reference microphone array alone.

The use of listener importance functions improves both the direction estimation and the position tracking accuracy (see Table 3.2, B $\in \{1^*, 2^*\}$). A succession of importance functions obtained from different speakers forces the particles of a listener to cumulate at the intersection points of the importance functions. One of the intersection points lies at or near the true location of the listener, thus allowing a rough estimate of the listener location from the particle locations (see Fig. 3.6, 20–30 s: tracking for the silent P2 re-converges to the true location around 28 s).

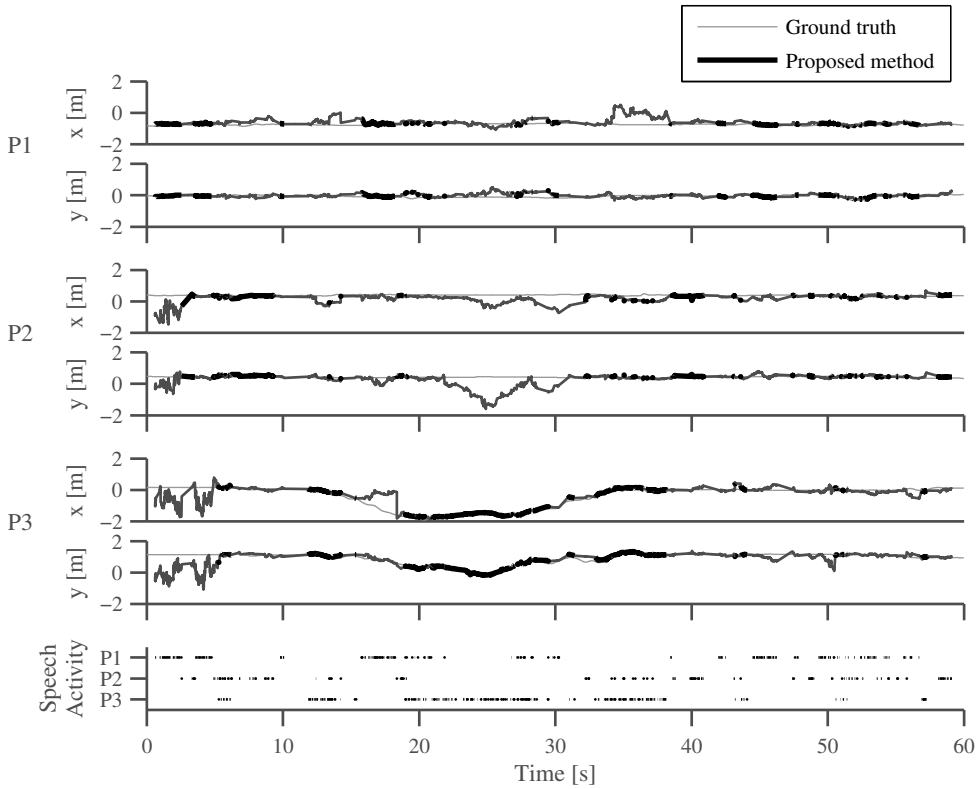

Figure 3.6. Tracking results and speech activity map. The active speaker, marked with a dot for each frame in the activity map, was detected as the conferee with the maximum headset microphone energy and sum of non-normalised particle weights. Speech activity was detected in 38% of all frames. The tracking results during frames where a conferee was active are marked as bold lines.

Best performance for the position tracking is achieved when using two binaural headset microphones and listener importance functions for each conferee. The tracking performance for this condition is shown in Fig. 3.6. With each conferee wearing just one microphone, the performance deteriorates slightly (see Table 3.2, B = 1*). The tracking RMSE of the proposed framework is comparable to values reported for state-of-the-art tracking methods under similar experimental conditions. Ward et al. (2003), Fallon and Godsill (2010), and Talantzis (2010) proposed the use of particle filtering to track acoustic sources in a room via microphone pairs delimiting the tracking area. The results are summarised in Table 3.3.

3.6.2 Orientation tracking

Figure 3.7 illustrates the tracking results for each conferee and the speech activity map. Speech activity was detected in 67% of the frames. The RMSE of the orientation tracking is given in Table 3.4. It is calculated for each conferee over the frames during which tracking was performed.

As a reference, the orientation tracking method proposed by Tikander et al.

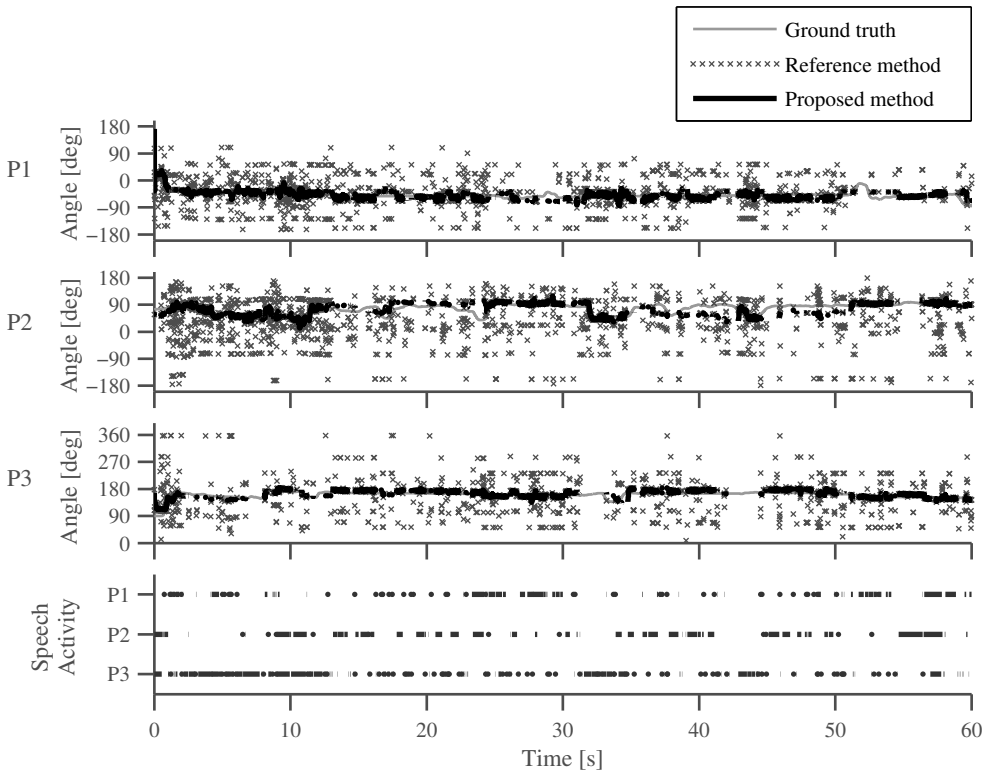

Figure 3.7. Head orientation tracking results for each of the three conferees P1, P2 and P3. The bottom graph indicates the frames were speech activity was detected.

(2004) was used. The reference method estimates the head orientation from a TDOA estimate between the binaural microphone signals of the listener, using a TDOA model:

$$\tau_{\mathrm{l,r}} = a\left(\varphi + \sin\varphi\right), \tag{3.28}$$

where a denotes the head radius.

For all three conferees, the proposed method clearly outperforms the reference method. This is partly due to the fact that the reference method estimates the head orientation based only on the TDOA estimate between the binaural microphone signals of the listener, whereas the proposed method uses also the TDOA estimates between the binaural microphones of the speaker and the listener. Furthermore, the fact that the particle filter takes into account past and current localisation information, through the history of each particle, seems to improve the tracking performance.

As can be seen in Fig. 3.7, P2 rotated the head the most during the meeting scenario. The RMSE is largest for P2, since a moving target generally suffers from a larger tracking error than a steady one. The tracking deteriorates in passages with large head movements or low speech activity, for instance around 15 s into the recording for P2. A key factor for the tracking performance is the SNR, calculated as the difference in dB between the signal energy and the noise floor. Fig. 3.8 illustrates the RMSE of both the proposed and the

	Orientation RMSE [deg]	
Conferee	(Tikander et al., 2004)	**here**
P1	28.97	**9.26**
P2	44.11	**11.95**
P3	30.39	**8.92**

Table 3.4. RMSE of the head orientation tracking for the reference method and the proposed approach. The results are calculated over the frames where tracking was performed.

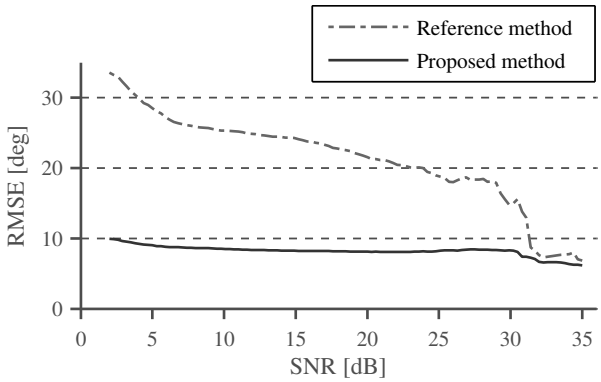

Figure 3.8. Orientation tracking RMSE versus SNR. The RMSE is averaged over three conferees.

reference method as a function of the SNR. The RMSE for each SNR value is obtained by averaging the RMSE of all three conferees over all frames with at least that SNR. As expected, the performance of both methods is better in frames with high SNR. Above 30 dB SNR the performance of the reference method approaches the performance of the proposed method. This implies that with high SNR a single TDOA estimate between the binaural microphones of the listener provides a reliable estimation of the head orientation, whereas the use of additional TDOA estimates in the proposed method yields only a minor improvement. In frames with low SNR, however, the proposed method clearly outperforms the reference method. Frames with low SNR provide weak evidence for tracking, hence in those frames the reference method fails, as it estimates the head orientation in each frame separately. The proposed method compensates for weak evidence in frames with low SNR by taking into account the tracking history, thus relying on strong tracking evidence found in frames with high SNR. Furthermore, the use of several TDOA estimates adds to the robustness of the proposed method.

3.7 Discussion

Methods for tracking user motion in an AAR environment are proposed in Publications I and II. The methods rely on microphone signals obtained from microphones embedded into the binaural ARA headset shown in Fig. 2.2. The tracking is based on time-delay estimation between microphone pairs and particle filtering.

For the speaker position tracking, a reference microphone array is combined with user-worn binaural headset microphones. The contribution of the location tracking method proposed in Publication II is twofold. Firstly, the improvement in tracking accuracy by employing user-worn microphones is shown. Secondly, a prior weighting method of the particles of silent conferees (i.e., the listeners) is proposed. It is based on deriving an importance function from the distance of each listener to the active speaker estimated from the signals of user-worn headset microphones. In an experimental setup, the locations of three conferees (two seated, one moving) engaged in a lively discussion were tracked. The root-mean square error (RMSE) for the speaker tracking was about 0.11 m using two binaural headset microphones per conferee, and about 0.13 m using one binaural headset microphone per conferee, which is comparable to the performance of state-of-the art acoustic tracking methods (see Table 3.3). The tracking performance obtained with just one user-worn microphone suggests that the proposed method may be suitable for other forms of user-worn microphones, including clip-on microphones attached to the clothing. The proposed listener importance function for prior particle weighting of the inactive conferees led to equal or improved tracking performance. Speaker tracking without user-worn microphones resulted in an RMSE of several metres, mainly due to speaker distance estimation errors, indicating a substantial improvement in tracking accuracy through the usage of user-worn microphones. The results show the proposed methods for speaker tracking and prior weighting of particles to be reasonably robust and accurate.

The orientation tracking method proposed in Publication I relies on the signals of binaural headset microphones. Unlike previously proposed methods that rely on anchor sources at known positions, the tracking is performed directly with the users' speech signals. In an experimental setup, the head orientations of three conferees in a meeting scenario were tracked. The RMSE of the proposed method is about 10 degrees. Although the orientation tracking depends on prior knowledge of the user locations, the locations could be inferred via the speaker location tracking method proposed here.

Future work includes the integration of dynamic models into the tracking algorithms, such as the Langevin model for location tracking (Ward et al., 2003) or the Ornstein-Uhlenbeck process for head orientation tracking (Lacouture-Parodi and Habets, 2013), to improve blind tracking performance during frames without speech activity.

4. Sound sample detection and numerosity estimation

An augmented reality (AR) system delivers information to the user as virtual content overlaid onto the environment. Research on information visualisation deals with the question how data can be presented to the user effectively by means of a graphical display (Ware, 2012). In information visualisation, two basic tasks relevant in a variety of applications are (i) detecting a certain element or sample among distractors, and (ii) estimating the percentage of certain elements or samples among distractors (Treisman, 1986; Julesz and Bergen, 1987; Healey et al., 1996; Michalski and Grobelny, 2008).

Publication III reports a user study investigating the performance of auditory display in these basic tasks adapted from information visualisation research. In the study, users were presented with lists of short sound samples, and asked to perform two tasks. Task I of the user experiment consisted in detecting a specific sound sample, referred to in this paper as the <key> sample, among distractor samples. In a practical application, the detection rate of a <key> sample is relevant when presenting points of interest in an AR navigation system, for example. In human vision, target elements can be detected and localised simultaneously (Sagi and Julesz, 1985). To test the hypothesis that detection and localisation can be done in parallel using auditory display, users were asked to state the perceived direction of the <key> sample. Task II of the user study investigated the user performance in estimating sample numerosity. Test participants were presented with two lists of short sound samples, and asked to determine which list contained more instances of the <key> sample. The ability to judge numerosity is relevant in an application conveying a general overview or "vibe" of an environment to the user (McGookin and Brewster, 2012).

The goal of the listening test was to study the effect of various auditory display design parameters on user performance in these two tasks. The parameters studied were derived from related work on auditory display.

4.1 Related work

To display information via auditory display in AAR, the information needs to be encoded as acoustic signals, if the primary source is not acoustic.

One option to automatically encode text for auditory display is via *text-to-speech synthesis*. The advantage of speech output is that the information presented can be readily interpreted by the users, without prior learning. Speech is used in public announcement systems, or in screen reader applications that allow visually impaired users access to verbal information (McGookin and Brewster, 2004a; Nees and Walker, 2009). On the downside, displaying information as speech can be slow, due to the sequential nature of speech (Sawhney and Schmandt, 2000). For non-verbal information, users may prefer non-speech sounds. Comparing speech and non-speech sounds in a navigation task, Tran et al. (2000) state that users found non-speech sounds easier to localise and more pleasant. *Sonification* refers to the process of mapping data to acoustic parameters of non-speech sound (Peres et al., 2008; Walker and Nees, 2011), e.g., in the form of auditory graphs (Brown et al., 2002; Nees and Walker, 2009; Batterman and Walker, 2013). *Auditory icons*, the acoustic counterpart of visual icons, employ metaphors to map sounds to their virtual referents (Gaver, 1986). Therefore, auditory icons are useful only if an intuitive mapping exists between information to be displayed and a sound. As an alternative to auditory icons, Blattner et al. (1989) introduced *earcons*. Earcons are abstract non-speech sounds that can be mapped to any item or process (Nees and Walker, 2009). Information is typically encoded in the form of a tone or short melody played by a musical instrument (Brewster et al., 1995b). Earcons provide the ability to convey hierarchical relationships through sound parameters, including rhythm, timbre, or pitch (Brewster et al., 1995a). Due to the abstract nature of earcons, the user needs to learn the association between an earcon and the information it represents (Garzonis et al., 2009). To minimise the learning required when using earcons, Walker et al. (2006) introduced *spearcons*, i.e., speech-based earcons. Spearcons are created by speeding up synthesised speech samples of the information to be displayed. Walker et al. (2013) studied the performance of spearcons for navigating auditory menus.

In the study reported in Publication III, earcons and synthesised speech were compared as two established and actively researched audio encoding strategies. Earcons were chosen as a non-speech alternative to synthesised speech. The specific characteristics of earcons as an encoding strategy for auditory display have been studied extensively elsewhere (Brewster et al., 1993,

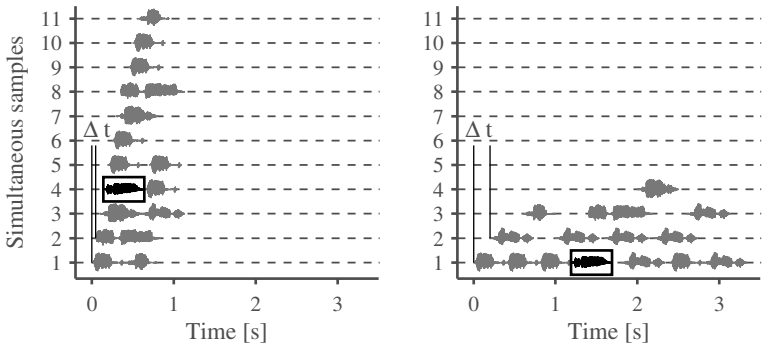

Figure 4.1. Audio sample sets containing 15 samples, staggered with (a) 50 ms and (b) 200 ms stimulus onset asynchrony (SOA), Δt. The <key> sample is highlighted.

1995b; McGookin and Brewster, 2004a), and are not considered here. Unlike prior work that has investigated a combination of earcons and speech in a user interface (Karshmer et al., 1994; Ramloll et al., 2001; Vargas and Anderson, 2003; Walker et al., 2006), the effectiveness of synthesised speech and earcons was compared separately. While earcons require learning (Dingler et al., 2008; Walker et al., 2013), adequate practice has been shown to lead to performance comparable to synthesised speech output in a dual attention task (Bonebright and Nees, 2009). To minimise the effects of learnability and memory on user performance, each participant of the study reported in Publication III had to concentrate on just one speech and one earcon sample representing the <key> sample, throughout the whole test.

When presenting a list of samples to the user, they have to be arranged in time. Earlier work has shown user performance to deteriorate as the number of maskers or distractors played concurrently with a <key> sample increases (Brungart et al., 2001; Brungart et al., 2002; McGookin and Brewster, 2004b). To study the effect of temporal overlap on user performance, the samples were displayed as a list staggered with a stimulus onset asynchrony (SOA). Based on findings by McGookin and Brewster (2004a) and a pilot study, a range of SOAs critical for user performance was determined and tested in the study. Figure 4.1 illustrates to sample sets with differing SOAs, Δt. As can be seen, an SOA of 50 ms results in up to eleven samples being played back concurrently, while at most four samples are presented simultaneously with an SOA of 200 ms.

Related to the temporal presentation is the spatial arrangement of the sound samples. Earlier work on spatial release from masking (SRM) and the "cocktail party effect" (see Section 2.2) indicates that spatial separation of concurrent sounds improves user performance (Bronkhorst, 2000; Brungart and Simp-

son, 2002; McGookin and Brewster, 2004a; Ihlefeld and Shinn-Cunningham, 2008a,b). However, masker type, spatial configuration, prior information on the target direction (i.e., "knowing where to listen"), and level differences between the target and maskers may affect user performance (Bregman, 1990; Darwin and Hukin, 2000; Brungart et al., 2001; Brungart et al., 2002; Kidd et al., 2005; Kidd et al., 2010). To investigate the effect of spatial separation on user performance, the study reported in Publication III compared spatial and non-spatial presentation of sample sets.

Prior research related to the tasks presented in Publication III studied the concept of "change deafness", i.e., the inability of the auditory system to detect changes in complex auditory scenes (Eramudugolla et al., 2005). In studies by Eramudugolla et al. (2005) and Pavani and Turatto (2008), test participants were asked to compare two auditory scenes that were identical except for the presence or location of a <key> element. Both studies found that the test subjects had difficulties perceiving changes in scenes containing between three and eight elements. However, the ability to perceive changes improved substantially when the test subjects' attention was directed to a specific <key> element. Both experiments reported in Publication III consisted of directed attention tasks. While the study by Eramudugolla et al. (2005) tested the test subjects' ability to detect object disappearance or a change in location when comparing two scenes, Task I in Publication III investigated the ability to detect object presence in a single scene in each trial. The study by Pavani and Turatto (2008) used animal calls to study the ability to detect object appearance or disappearance when comparing two scenes. Both tasks in Publication III used either synthesised speech or earcons as examples of well-established sound types for auditory display.

4.2 Experimental design and procedure

A listening test was conducted to investigate the effects of various auditory display design parameters on user error rates in two tasks adapted from information visualisation: (i) detecting the <key> sample, and (ii) estimating sample numerosity. The independent variables for the study were derived from auditory display design parameters: (i) the audio encoding strategy, (ii) the temporal arrangement, and (iii) the spatial arrangement. For the numerosity estimation, a fourth independent variable was the relative numerosity of the <key> sample in two sets. A schematic overview of the experimental design is shown in Fig. 4.2. The audio encoding strategies compared in the study were

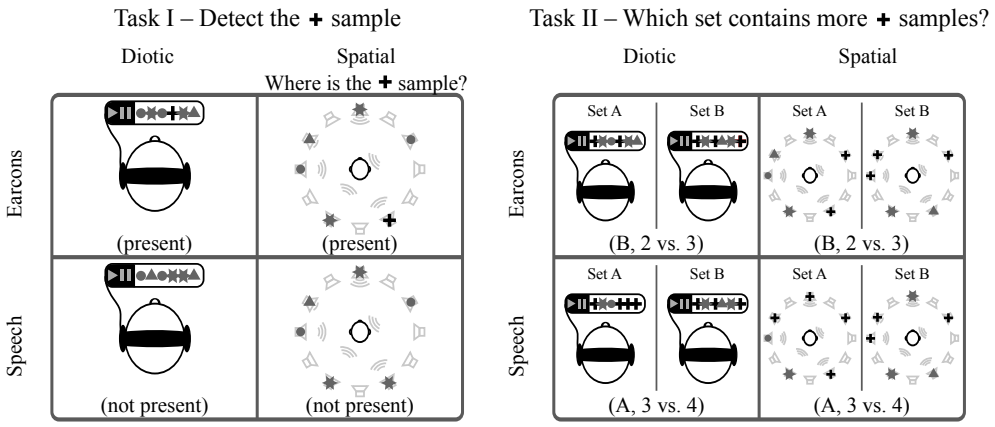

Figure 4.2. Experimental setup. The cross symbol represents the <key> sample; all other symbols represent distractors. In task I, each set contained a total of 15 samples staggered with an SOA of 50, 100, 200, or 400 ms. In task II, each set contained two to seven instances of the <key> sample, totalling 10 or 20 samples staggered with an SOA of 100, 200, or 400 ms.

synthesised speech and earcons. For the temporal arrangement, SOAs ranging from 50 ms to 400 ms were tested. To present the samples spatially separated, a multichannel loudspeaker system was used. Non-spatial presentation was implemented via diotic headphones-playback. The sample numerosities tested in the numerosity estimation task were 3 vs. 4, 2 vs. 3, 3 vs. 5, 3 vs. 6, and 3 vs. 7.

4.2.1 Test conditions

In task I, test subjects were presented with sets of 15 sound samples. For each set, the subjects were asked to determine whether the <key> sample was present. For the spatial presentation, the subjects were asked to indicate from which loudspeaker the <key> sample was presented. The test hypothesis was that the subjects would be able to recall the direction of the <key> sample. In task II, test subjects were presented with two sets of 10 or 20 samples, containing two to seven instances of the <key> sample each. The subjects had to determine whether the two sets contained the same number of <key> samples, or which set contained more. The test hypothesis was that larger relative numerosity differences would be easier to detect than small differences.

In both tasks, the sound samples were staggered with an SOA ranging from 50 ms to 400 ms, the hypothesis being that a larger SOA would improve user performance by decreasing the temporal overlap between samples.

The sound samples were presented either diotically via headphones or with randomised directions via a multi-channel loudspeaker setup. For the diotic playback, the anechoic monophonic input signal was presented to both ears, allowing precise control over the ear input signals and minimising the effect

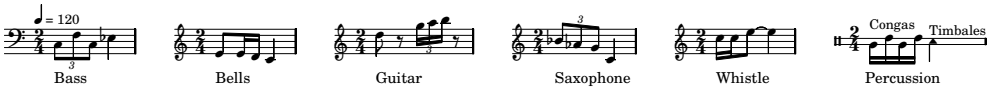

Figure 4.3. Earcons used in the user study. The timbres were produced via Apple OSX GarageBand's inbuilt MIDI instruments.

of head rotation. The loudspeaker playback ensured accurate reproduction of localisation cues, thus maximising the potential benefit of displaying the samples spatially separated along a circle in the horizontal plane.

4.2.2 Apparatus and sound samples

The study was conducted in the same space as the experimental evaluations of the motion tracking algorithms (see Section 3.5) proposed in Publications I and II. The speech samples used in the study were obtained by synthesising the words "book", "chair", "keys", "microwave", "couch", and "cup" via the Apple OSX's inbuilt speech synthesiser. Correspondingly, six earcons were generated using Apple OSX's GarageBand via the inbuilt MIDI instruments with the following timbres: "Bass", "Bells", "Guitar", "Saxophone", "Whistle", and "Percussion". A transcription of the earcons is shown in Fig. 4.3. To ensure equal loudness, all samples were normalised using A-weighting.

4.2.3 Test procedure

The study reported in Publication III was laid out in a fully randomised, within-subject design. None of the 22 test participants reported any hearing impairments. The duration of the experiment was about 90 minutes. The participants were seated in the centre of the circular loudspeaker setup and asked to report their answers to the listening tasks in a questionnaire. The <key> samples, one earcon and one synthesised speech sample per test subject, were introduced at the beginning of the experiment.

4.3 Results

To evaluate the results of the listening test, the total error count was calculated for each tested condition in both tasks and summarised in a contingency table. The reported p-values were obtained via pairwise Pearson's chi-squared tests. For multiple comparisons, a Holm-Bonferroni correction is applied to the p-values (Holm, 1979; Dudoit et al., 2003).

	SOA [ms]				**Playback**		**Sound type**	
	50	100	200	400	Diotic	Spatial	Earcon	Speech
Trials	440	440	440	440	880	880	880	880
Errors	104	46	26	11	80	107	73	114
Errors [%]	24	10	6	2	9	12	8	13
p-value	<0.001		0.037	0.037	0.044		0.002	

Table 4.1. Error count table for Task I. Chi-squared tests indicate statistically significant differences between all adjacent columns of the independent variables.

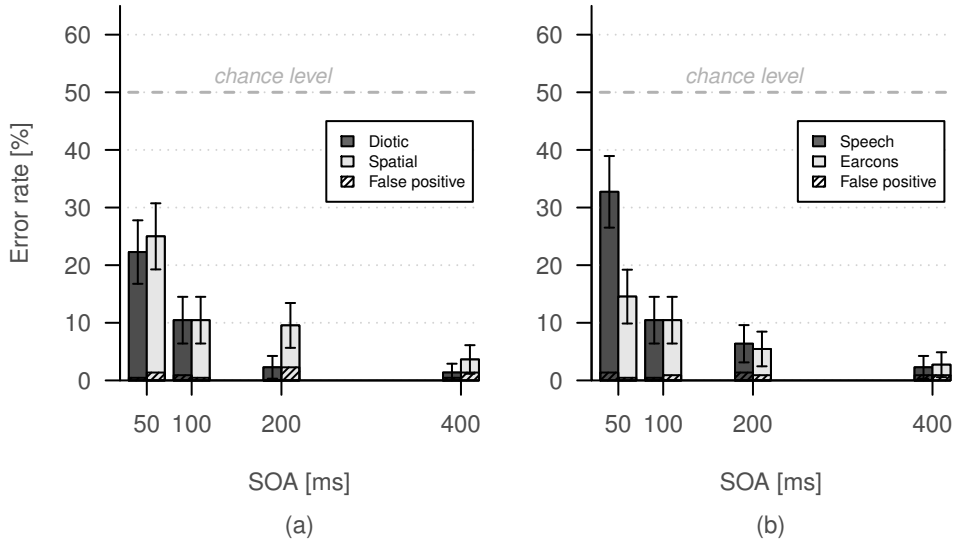

Figure 4.4. Average error rates as a function of (a) playback condition and (b) sound type for different SOAs. The hatched area indicates false positive errors. Error bars indicate 95% confidence intervals for the means.

4.3.1 Task I: detect the <key> sample

In task I, 80% of all trials contained the <key> sample. Test subjects were not aware of this distribution. Therefore, guessing the presence or absence of the <key> sample would yield an error rate of 50%.

The results indicate that error rates decrease for each SOA increase, the effect being statistically significant. No substantial difference in user performance was found between spatial playback via the multichannel loudspeaker setup and diotic headphone playback. Earcons slightly outperformed synthesised speech, with the "bass" earcon (see Fig. 4.3) being correctly identified in all trials. The results are summarised in Table 4.1.

Figure 4.4 shows the average error rates of (a) both playback conditions and (b) both sound types across the tested SOAs. False positive errors, i.e., a user indicating that the <key> sample was present in the set when it was not, account for less than 9% of the total error rate (see Fig. 4.4, hatched area).

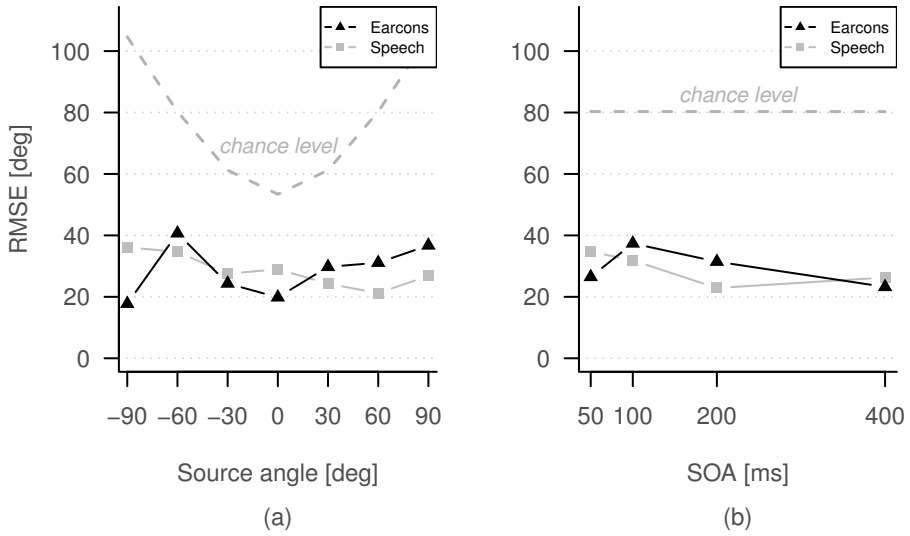

Figure 4.5. Angle RMSE of speech and earcon playback as a function of (a) the lateral angle of the <key> sample and (b) the SOA.

For the spatial loudspeaker playback, the test subjects were asked to state from which loudspeaker they thought the <key> sample was being played. Figure 4.5 shows the root-mean square error (RMSE) of the direction estimates, calculated from the difference between the actual and the perceived lateral angle of a correctly detected <key> sample. A visual inspection of the graphs in Fig. 4.5 indicates that the RMSE does not differ substantially between earcon and speech playback, and that there is no visible dependency from (a) the lateral angle of the <key> sample or (b) the SOA.

4.3.2 Task II: estimate the <key> sample numerosity

In task II, the distribution of <key> samples was randomised, with set A containing more instances of the <key> sample in 50% of the cases and set B containing more instances in the remaining 50% of the cases. Test subjects were not aware of the sample distribution. Therefore, guessing would result in a 67% error rate.

As in task I, there is a statistically significant decrease of the error rates for each increase of the SOA. A similar relationship holds between error rates and the numerosity of the <key> samples. Here, the numerosity is expressed as a relative difference in per cent. For example, 3 <key> samples in set A vs. 4 <key> samples in set B corresponds to a relative difference of 33%. Error rates decreases statistically significantly for each increase of the relative numerosity difference. The error rates under both playback conditions, i.e., diotic headphone and spatial loudspeaker playback, are equal. Synthesised

	SOA [ms]			**Difference [%]**				
				33	50	67	100	133
	100	200	400	(3 vs. 4)	(2 vs. 3)	(3 vs. 5)	(3 vs. 6)	(3 vs. 7)
Trials	880	880	880	528	528	528	528	528
Errors	433	238	103	254	215	131	103	71
Errors [%]	49	27	12	48	41	25	20	13
p-value	<0.001	<0.001		0.037	<0.001	0.045	0.030	

	Playback		**Sound type**	
	Diotic	Spatial	Earcon	Speech
Trials	1320	1320	1320	1320
Errors	385	389	412	362
Errors [%]	29	29	31	27
p-value	0.898		0.036	

Table 4.2. Error count table for Task II. Chi-squared tests indicate statistically significant differences between all adjacent columns of the independent variables, except for "playback type", i.e., diotic headphone and spatial loudspeaker playback.

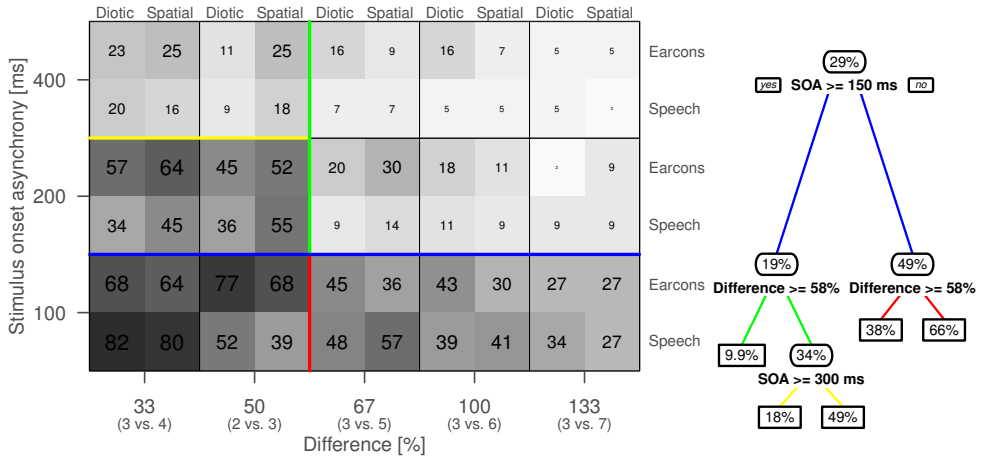

Figure 4.6. (Left) error rates (in per cent) as a function of SOA, difference between the number of <key> samples, playback condition (diotic headphone or spatial loudspeaker playback), and sound type (speech or earcons). The <key> sample numerosity is expressed as a relative difference in per cent. Pure guessing would give an error rate of 67%. (Right) a regression tree of the user performance. The nodes indicate average error rates.

speech playback slightly outperforms earcons. The results are summarised in Table 4.2.

Figure 4.6 displays (left) the error rates for all levels of the independent variables and (right) a regression tree obtained via the R rpart library (Therneau et al., 2011). Both the regression tree analysis and a visual inspection of the plot confirm the SOA and relative difference to be the main determinants affecting user performance. Furthermore, a strong interaction between the SOA and the

relative difference is visible. For small relative differences (33% and 50%) and an SOA of 100 ms, the average error rate is 66%, which corresponds to random chance (Fig. 4.6, bottom left). A larger relative difference or SOA improves performance substantially. With an SOA of at least 200 ms and a relative difference of at least 67%, the average error rate drops below 10% (Fig. 4.6 (left), top right).

4.4 Discussion

The user study presented in Publication III investigates the effect of auditory display parameters on user performance for detecting a <key> sample and estimating its numerosity. As expected, error rates in both tasks decreased significantly with each SOA increase, due to the reduced temporal overlap of samples staggered with a larger SOA. However, with an SOA of at least 100 ms, error rates for detecting a <key> sample dropped to about 10%, indicating that auditory display is effective for sample detection tasks even with a dense temporal sample arrangement. This result is largely in line with findings by Pavani and Turatto (2008), who reported an average error rate of 18% for detecting the presence of a <key> element in auditory scenes consisting of animal calls.

Contrary to the test hypothesis, spatial separation of the samples did not improve user performance in either task. This may be explained by the fact that the directions of the <key> samples were randomised and thus unknown to the listener beforehand. The lack of a priori information about the target direction may have cancelled the advantage of spatial separation, as indicated by earlier studies (Brungart et al., 2002; Kidd et al., 2005). As a consequence, in practical applications, diotic or indeed monophonic playback may be sufficient to convey the presence or numerosity of a target sample. However, for spatially separated samples, users were able to estimate the direction of a <key> sample, once detected, relatively accurately.

The results suggest that earcons and synthesised speech were similarly effective for the detection and numerosity estimation tasks. While the samples used in the study could be further optimised to improve performance, the finding that samples obtained via text-to-speech synthesis performed similarly to non-speech sounds may be of interest for practical applications that need the encoding of data into sound to be automated.

5. Rendering virtual sources

AAR applications rely on the ability to render virtual sound sources at an arbitrary direction or position in space. This requires encoding appropriate localisation cues into the ear input signals. For a real source in free field, these localisation cues are described by the head-related transfer function (HRTF) (see Section 2.4.5). When displaying a virtual source over headphones, for instance in a mobile AAR application, the localisation cues can be encoded by filtering the sound signal with an appropriate pair of left and right HRTFs (see Section 2.4.5). This allows rendering spatialised sound via a pair of headphones, for instance in mobile AAR applications.

There are, however, a few caveats when rendering a virtual source over headphones using HRTFs. As discussed in Section 2.5.7, despite encoding appropriate localisation cues, the virtual source may be perceived inside the head rather than externalised. Steps to improve the externalisation of virtual sources include the careful calibration of the playback system, the usage of individually measured HRTFs, and the inclusion of measured or artificial reverberation. Furthermore, a virtual source in the front may be perceived as emanating from the back and vice versa, a problem referred to as front–back confusion. To prevent front–back confusions, and to keep a virtual source correctly registered with the real environment in the presence of head movements, dynamic cues need to be rendered.

To render dynamic virtual sources with high fidelity in a mobile AAR application, the spatial rendering system needs to reproduce spatial cues accurately, to avoid inside-the-head locatedness (IHL), render dynamic cues, to support interaction and avoid front–back confusions, and run in real time. Furthermore, it may be desirable to minimise computational load, to prolong battery life, and audible artefacts caused by the rendering system. Such artefacts can occur when the filters used to encode spatial cues in the ear signals change abruptly. To improve the accuracy and smoothness of the spatial rendering process,

HRTF interpolation can be employed. In Publications IV and V, an HRTF interpolation framework is proposed that has low computational complexity and can be applied to HRTF measurement sets with arbitrary measurement positions, including measurements taken at various distances.

5.1 Head-related transfer function interpolation

When rendering virtual sources for AAR applications, the source direction or position relative to the listener changes if either the listener or the source moves. To account for this change, the rendering system needs to update the rendering filters that encode the spatial cues accordingly. The rendering filters are often obtained from measured HRTF datasets. As HRTFs are typically measured on a discrete grid, a straightforward way to update the spatial filters is to use the HRTFs closest to the desired direction or position of the virtual source. However, if the desired source direction or position does not fall on an HRTF measurement point, switching to the closest measured HRTFs may result in an audible spectral change of the ear input signals (Zotkin et al., 2004). Furthermore, for a coarse HRTF measurement grid, switching to the closest measured HRTF would give incorrect spatial cues and introduce localisation errors.

To prevent audible artefacts due to the filter update, and to provide more accurate spatial cues, intermediate HRTFs can be estimated via interpolation from HRTFs measured on a discrete grid. Experiments by Langendijk and Bronkhorst (2000) have shown that HRTF interpolation successfully restores spatial cues in virtual sources rendered at directions not available in the HRTF dataset, if the spatial resolution of the dataset is about 6 degrees.

HRTFs are typically measured at a fixed distance from the test subject, over a range of azimuth and elevation angles spaced at regular intervals. Examples of publicly available HRTF databases include the MIT KEMAR database (Gardner and Martin, 1995), the CIPIC database (Algazi et al., 2001a), and the LISTEN database (IRCAM, 2013). When using one of these databases for spatial rendering, a straightforward way to arrive at an intermediate HRTF estimates is to take a weighted average of neighbouring HRTF measurements via linear interpolation. More complex interpolation approaches take into account more or all measurements, for example by using spherical splines (Hartung et al., 1999). The interpolation itself may be performed in the time domain directly on the delay-aligned HRTFs (Hartung et al., 1999) or using a minimum-phase representation (Wenzel and Foster, 1993), in the frequency domain on

the magnitude spectra (Zotkin et al., 2004; Carty and Lazzarini, 2009), or using other representations including principal components (Wang et al., 2009) or spherical harmonics (Zotkin et al., 2009).

While the more sophisticated approaches to HRTF interpolation potentially yield smoother and more accurate HRTF estimates (Luo et al., 2013), they come at the cost of increased complexity both in terms of implementation and computation. Therefore, spatial rendering approaches for practical real-time systems, including AAR applications, typically rely on the simpler and computationally less demanding linear interpolation of nearest neighbours (Savioja et al., 1999; Freeland et al., 2002; Queiroz and Sousa, 2011). The steps for rendering a virtual sound source using linear HRTF interpolation are

1. select neighbouring HRTF measurements;
2. calculate interpolation weights;
3. interpolate HRTFs;
4. filter the input signal with the interpolated HRTFs.

The neighbour selection and calculation of interpolation weights is typically done based on some distance criterion. The interpolation can be performed for instance in the time or frequency domain, as discussed above. The filtering of the input signal is typically implemented via fast block-convolution; after employing the Fast Fourier Transform (FFT), blocks of the input signal are convolved with the spatial filters in the frequency domain via overlap-add or overlap-save (Oppenheim et al., 1999; Välimäki et al., 2012). To allow dynamic rendering using head tracking, the block size should be chosen small enough to allow a sufficient update rate (see Section 2.5.6).

While several studies in the literature have investigated ways to improve and optimise steps 3 and 4 in the above list, relatively little attention has been given to the way in which a subset of neighbouring HRTFs is selected and the interpolation weights are calculated. Furthermore, most prior work has focussed on 2-D HRTF interpolation, i.e., interpolation in azimuth and elevation. Next, previously proposed approaches to subset selection and the calculation of interpolation weights for 2-D interpolation are discussed.

5.1.1 Subset selection

Proposed methods for selecting a subset of HRTF measurements for linear interpolation include finding the nearest three measurement points (Jot et al., 1995; Zotkin et al., 2004) and finding the nearest four measurement points (Wenzel

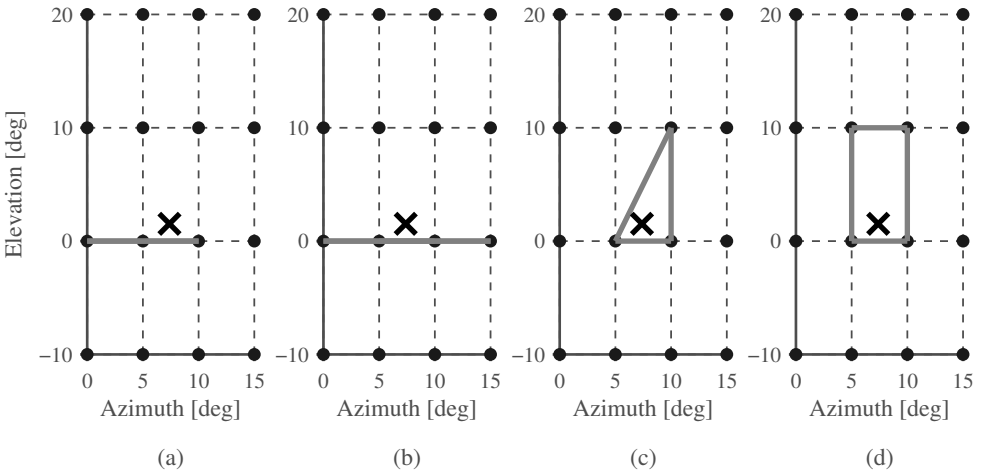

Figure 5.1. Subset selection approaches: (a) nearest-3, (b) nearest-4, (c) enclosing triangle, and (d) enclosing rectangle. The dots denote measurement points in the MIT KEMAR HRTF database (Gardner and Martin, 1995), the cross denotes the desired source direction.

and Foster, 1993; Hartung et al., 1999). Distance measures proposed to determine the nearest neighbours include the great-circle distance (Hartung et al., 1999) and the Euclidean distance (Zotkin et al., 2004). Another subset selection approach is to select the nearest measurement points that enclose the desired source direction. Proposed approaches include finding a rectangle (Savioja et al., 1999; Langendijk and Bronkhorst, 2000) or triangle (Freeland et al., 2004; Queiroz and Sousa, 2011) enclosing the desired direction.

Figure 5.1 shows an example of the various subset selection approaches. With the approaches depicted in Fig. 5.1a and 5.1b, a subset of HRTF measurements is selected for interpolation based on the distance to the desired virtual source direction. The advantage of a simple distance criterion for selection is that it works regardless of the measurement grid layout. However, it may result in selecting HRTFs that do not enclose the desired source direction, and instead lie on a line (see Fig. 5.1a and 5.1b).

Figure 5.1c and 5.1d illustrate subset selection based on the criterion that the selected HRTFs enclose the desired source direction. While this approach is well-suited for linear interpolation, it can not be implemented in a straightforward way if the HRTF measurement grid exhibits irregularities. When measuring HRTFs on a spherical grid, the measurement points are typically distributed more sparsely towards the poles than at the equator, in accordance with the decreasing localisation accuracy of humans towards extreme elevations (Gardner and Martin, 1995; IRCAM, 2013). Other potential causes for irregularities in the HRTF measurement grid are movements of human subjects during HRTF measurements (Bolaños and Pulkki, 2012), and positioning errors of the

mechanical measurement setup.

5.1.2 Calculation of interpolation weights

Bilinear interpolation can be used to calculate weights to interpolate measured HRTFs linearly with respect to azimuth and elevation. Bilinear interpolation has been proposed for interpolating three measurement points (Freeland et al., 2004), and four measurement points arranged in a regular grid (Savioja et al., 1999). Geometric approaches calculate weights based on the distance of the desired direction from each measurement point. For the interpolation of three or more measurement points, weights can be calculated from the inverse of the Euclidean distance (Zotkin et al., 2004) or the great-circle distance (i.e., the distance along the sphere) (Hartung et al., 1999; Carlile et al., 2000). The interpolation of measured HRTFs can be interpreted as a superposition of the signals of virtual loudspeakers positioned at the measurement points (Queiroz and Sousa, 2011). With this interpretation, the interpolation weights are analogous to the panning gains of these virtual loudspeakers. Panning gains for arbitrary loudspeaker setups can be calculated using VBAP (Pulkki, 1997).

5.1.3 Interpolation in azimuth, elevation, and distance

While far-field HRTFs can be considered distance-independent (Brungart, 2002; Kan et al., 2009), the faithful rendering of virtual sources in the near-field requires the use of distance-dependent HRTFs (Brungart et al., 2001). A number of approaches have been proposed previously to estimate near-field HRTFs from HRTFs measured at a fixed distance (Duraiswami et al., 2004; Menzies and Al-Akaidi, 2007; Romblom and Cook, 2008; Kan et al., 2009; Spors and Ahrens, 2011). However, recently published HRTF databases containing measurements obtained at various distances in the near-field (Qu et al., 2009; Yu et al., 2010; Bolaños and Pulkki, 2012) call for 3-D interpolation methods. Previously proposed methods to interpolate near-field HRTFs include interpolating two HRTFs to estimate HRTFs at an intermediate distance (Lentz et al., 2006), or eight HRTFs forming a volume enclosing the desired source position (Villegas and Cohen, 2010). However, these previously proposed 3-D interpolation approaches rely on ad hoc methods for subset selection and interpolation weight calculation tailored for a specific HRTF database, and can therefore not be directly applied to arbitrary HRTF measurement grid layouts. An example of an HRTF database with grid irregularities is shown in Fig. 5.3. The irregularities are mainly caused by movements of the test subjects

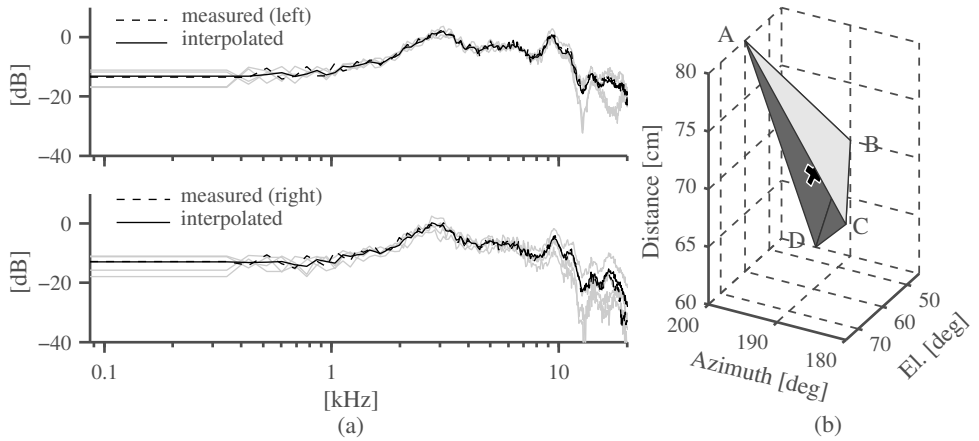

Figure 5.2. (a) Magnitude spectra of measured (dashed line) and estimated (solid line) HRTF via linear interpolation of measured HRTFs (light lines); (b) desired source position $\mathbf{x}$ and HRTF measurement positions forming tetrahedron used for interpolation. The measurements are taken from the database by Yu et al. (2010).

during the HRTF measurements. These movements are difficult to avoid when performing measurements on human subjects. Therefore, there is a need for an HRTF interpolation framework that can cope with grid irregularities inherent in datasets of human HRTFs.

5.2 Proposed approach

The proposed HRTF interpolation framework is based on linear interpolation of a minimal subset of measured HRTFs. It builds on prior work on HRTF interpolation methods (see Section 5.1), and presents a general interpolation framework applicable to HRTF measurement databases with arbitrary measurement grid layouts.

Given a set of HRTFs measured at a fixed distance, an interpolated HRTF can be obtained from three measurement points forming a triangle enclosing the desired source *direction* (Freeland et al., 2004; Queiroz and Sousa, 2011) (see Fig. 5.1c). The present work shows how this approach can be extended to include the desired source distance, through direct interpolation of HRTF measurements obtained at various distances that form a tetrahedron enclosing the desired source *position* (see Fig. 5.2b). To minimise computational load at run time, the HRTF measurements are grouped into subsets suitable for linear interpolation during initialisation. Next, this pre-processing of HRTF data is described.

5.2.1 Triangulation of measurement points

Linear HRTF interpolation is based on the assumption that an HRTF estimate for a desired source direction or position can be obtained by interpolating a subset of HRTF measurements close to the desired source. To avoid ambiguities when selecting the subset there should be a unique representation of an HRTF estimate for a specific direction or position as a linear combination of neighbouring HRTF measurements. To meet these requirements, the measurement points are grouped such that they form non-overlapping geometric simplices, i.e., triangles or tetrahedra. The grouping is done during initialisation, to minimise computational load during run time.

A set of points in 2-D can be grouped into non-overlapping triangles via *triangulation*. For a set of points lying on the surface of a sphere, taking the convex hull yields a Delaunay triangulation (Aurenhammer, 1991). When using triangles for interpolation, it is desirable that they be nearly equiangular. The Delaunay triangulation is optimal in this sense, and it maximises the minimum angle of the generated triangles (Aurenhammer, 1991).

Efficient algorithms exist to perform the Delaunay triangulation in 2-D and 3-D (Aurenhammer, 1991). For points lying on a plane, the Delaunay triangulation generates triangles such that the circumcircle of each triangle contains no other points (Aurenhammer, 1991). In 3-D, the Delaunay triangulation yields tetrahedra such that the circumsphere of each tetrahedron contains no other points. Figure 5.3a illustrates the Delaunay triangulation for 100 random measurement points distributed over the surface of a sphere, in analogy to HRTF measurements taken at random directions with fixed distance. As can be seen, the triangulation deals well with grid irregularities. Figure 5.3b and 5.3c illustrates the tetrahedral mesh generated via Delaunay triangulation of a set of 3-D measurement points, for (b) an HRTF database with a highly regular measurement grid, and (c) a database with grid irregularities. The mesh consists of non-overlapping tetrahedra that fill the space occupied by the measurement grid. Any point inside that space is enclosed by exactly one tetrahedron, except if the point lies on a vertex, edge, or facet shared by multiple tetrahedra.

5.2.2 Calculation of interpolation weights

Once a mesh of non-overlapping simplices has been generated from the HRTF measurement points via triangulation, an HRTF estimate for any desired source position, $\mathbf{x}$, lying inside the mesh can be obtained by interpolating the

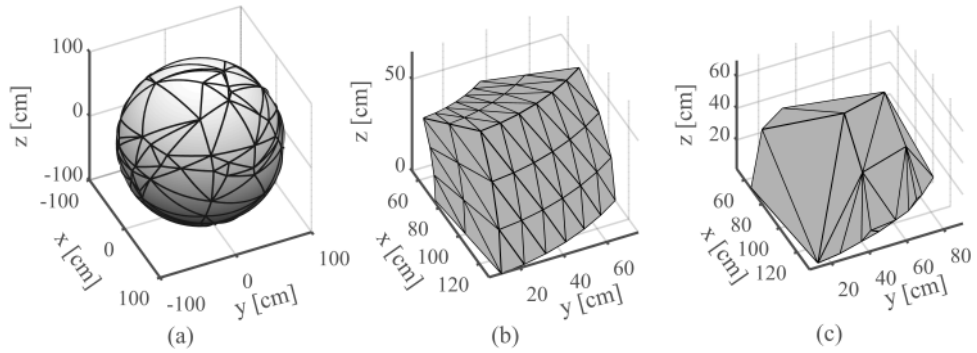

Figure 5.3. Delaunay triangulation of (a) 100 random HRTF measurement points at a fixed distance, (b) the PKU&IOA HRTF database (Qu et al., 2009), (c) the HRTF database by Bolaños and Pulkki (2012).

HRTFs forming the vertices of the simplex enclosing $\mathbf{x}$. Next, the calculation of interpolation weights is discussed for a tetrahedral mesh of HRTF measurements taken at various distances. The interpolation of triangles on the surface of a sphere can be interpreted as a special case of the tetrahedral interpolation.

Consider a tetrahedron formed by the vertices, $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$, as depicted in Fig. 5.2b. Any point, $\mathbf{x}$, inside the tetrahedron can be represented as a linear combination of the vertices:

$$\mathbf{x} = g_1\mathbf{A} + g_2\mathbf{B} + g_3\mathbf{C} + g_4\mathbf{D}, \tag{5.1}$$

where g_i are scalar weights. With the additional constraint

$$\sum_{i=1}^{4} g_i = 1, \tag{5.2}$$

the weights, g_i, are the *barycentric coordinates* of the point, $\mathbf{x}$ (Sundareswara and Schrater, 2003). The barycentric coordinates can directly be used as interpolation weights for estimating the HRTF, $\hat{\mathrm{H}}_\mathrm{x}$, at the point, $\mathbf{x}$, as the weighted sum of the HRTFs, H_i, measured at the vertices, $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$:

$$\hat{\mathrm{H}}_\mathrm{x} = \sum_{i=1}^{4} g_i \mathrm{H}_i. \tag{5.3}$$

Subtracting $\mathbf{D}$ from both sides of Eq. (5.1) yields

$$\mathbf{x} - \mathbf{D} = \begin{bmatrix} g_1 & g_2 & g_3 \end{bmatrix} \mathbf{T}, \tag{5.4}$$

where

$$\mathbf{T} = \begin{bmatrix} \mathbf{A} - \mathbf{D} \\ \mathbf{B} - \mathbf{D} \\ \mathbf{C} - \mathbf{D} \end{bmatrix}. \tag{5.5}$$

Given a desired source position $\mathbf{x}$, the barycentric interpolation weights are found by evaluating

$$\begin{bmatrix} g_1 & g_2 & g_3 \end{bmatrix} = (\mathbf{x} - \mathbf{D})\,\mathbf{T}^{-1}, \tag{5.6}$$

and, with Eq. (5.2),

$$g_4 = 1 - g_1 - g_2 - g_3. \tag{5.7}$$

Note that $\mathbf{T}$ depends solely on the geometry of the tetrahedron and is independent of the desired source position $\mathbf{x}$. Therefore, $\mathbf{T}^{-1}$ can be pre-calculated for all tetrahedra during initialisation and stored in memory. This reduces the operational count for finding the interpolation weights via Eqs. (5.6) and (5.7) to twelve additions and nine multiplications per tetrahedron.

The interpolation of HRTFs measured at a fixed distance can be interpreted as a special case of the tetrahedral interpolation with a dummy vertex at the origin. Assume a triangular mesh obtained via a Delaunay triangulation from HRTF measurement points on the surface of a sphere. Given a triangle with vertices, $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and a dummy vertex at the origin, $\mathbf{D}$, a tetrahedron is formed with $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$. With $\mathbf{D} = \mathbf{0}$, i.e., the null vector, Eq. (5.5) and Eq. (5.6) simplify to

$$\mathbf{T} = \begin{bmatrix} \mathbf{A} \\ \mathbf{B} \\ \mathbf{C} \end{bmatrix} \tag{5.8}$$

and

$$\begin{bmatrix} g_1 & g_2 & g_3 \end{bmatrix} = \mathbf{x}\,\mathbf{T}^{-1}, \tag{5.9}$$

where g_1, g_2, g_3 are the barycentric interpolation weights for the HRTFs measured at the vertices of the triangle, $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$. Note that the source position, $\mathbf{x}$, obtained from the desired source azimuth and elevation, lies on the surface of the sphere, and thus outside the tetrahedron formed by the vertices, $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$. With this assumption, the weights calculated in Eq. (5.9) are analogous to the panning gains used in vector base amplitude panning (VBAP) (Pulkki, 1997).

Barycentric weights are well-suited for interpolation for a variety of reasons:

- For a point lying inside a triangle or tetrahedron, the barycentric weights g_i are positive: $0 < g_i < 1$.
- For a point moving inside a triangle or tetrahedron, the weights change smoothly as a function of the vertex-distance (Sundareswara and Schrater, 2003).
- For a point lying on a vertex $\mathbf{A}$, the barycentric weights are 1 at $\mathbf{A}$ and 0 otherwise, hence the interpolation at $\mathbf{A}$ is exact.
- For a point lying on an edge of a triangle, only the vertices forming that edge have nonzero barycentric weights. Furthermore, the vertex weights

for all triangles sharing that edge are identical. The same is true for a point lying on an edges or facets of a tetrahedron.

The above properties are particularly advantageous for the display of moving virtual sources, as the interpolation via barycentric weights does not cause discontinuities in the interpolated HRTFs. For a source moving smoothly from one triangle or tetrahedron to another across a shared vertex, edge, or facet, the HRTF estimate changes smoothly at the crossing point. Figure 5.2a illustrates the HRTF interpolation for a source position, $\mathbf{x}$, using barycentric weighting of the HRTF measurements at the vertices of the enclosing tetrahedron, depicted in Fig. 5.2b.

5.2.3 Selecting a subset for interpolation

Given the triangulation of HRTF measurements and a desired source position, $\mathbf{x}$, a suitable subset for interpolation can be found by evaluating the barycentric coordinates: $\mathbf{x}$ lies inside a simplex if and only if all barycentric coordinates are positive. Therefore, a straightforward way to find a suitable subset for interpolation is to iterate through the mesh until a simplex is found that satisfies this condition. An example run of this linear-time "brute-force" approach for a tetrahedral mesh is shown in Fig. 5.4a.

Given the large number of simplices generated for dense HRTF measurement grids, and the tight processing time constraints of real-time audio applications, it is desirable to speed up the process of selecting a subset for interpolation. A more efficient way to locate a point in a triangulation is via an *adjacency walk* (Sundareswara and Schrater, 2003). Starting from a random tetrahedron, evaluate the barycentric coordinates and walk to the adjacent tetrahedron across the triangle formed by the vertices with the three largest barycentric coordinates; terminate when all barycentric coordinates are positive (see Fig. 5.4b, light grey tetrahedra). The adjacency walk algorithm for a triangle mesh is analogous. The theoretical complexity of the adjacency walk for non-homogenous meshes is $O((n)^{\frac{1}{m}})$ (Sundareswara and Schrater, 2003), where n is the number and m the dimensionality of the vertices. This constitutes a substantial improvement in terms of scalability over the $O(n)$ brute-force approach. As shown in Fig. 5.4c, the worst-case performance of the adjacency walk (AW) is well below 0.1 ms even for the largest tested database (31752 tetrahedra). To reduce the number of steps needed for the adjacency walk to terminate in a tetrahedral mesh, a tetrahedron close to the desired source position $\mathbf{x}$ can be chosen as the starting point for the walk. A simple yet efficient way to find the closest neighbours to a

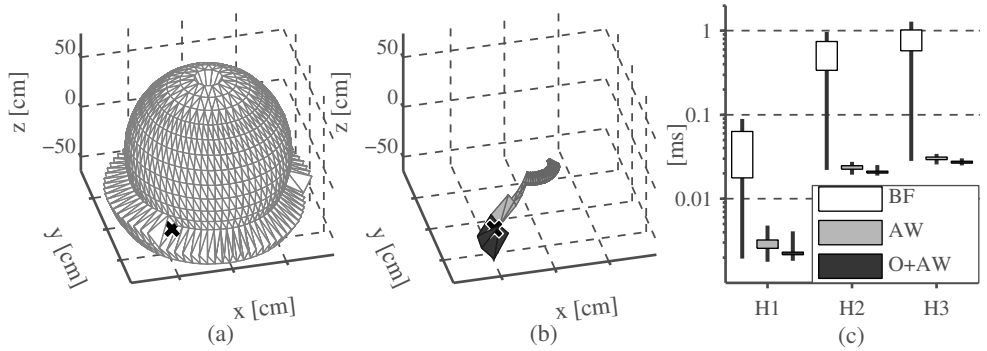

Figure 5.4. Tetrahedron selection for source position $\mathbf{x}$ via (a) brute-force (BF) search (18772 iterations) and (b) adjacency walk (AW) with a random starting tetrahedron (light grey, 105 iterations) and with a tetrahedron selected via an octree query (O+AW, dark grey, 5 iterations); (c) running-times for 1000 random source positions, averaged over 100 repetitions, on a computer with a 2GHz quad-core processor, for 3 HRTF databases: H1 (Bolaños and Pulkki, 2012), H2 (Yu et al., 2010), H3 (Qu et al., 2009); vertical lines extend from minimum to maximum, boxes from lower to upper quartile.

point in 3-D is by querying an *octree* representation of the HRTF measurement points (Samet, 1989). A cuboid containing all points forms the root of the octree. Starting from the root cuboid, the octree is generated by recursively dividing every cuboid into eight equal-sized cuboids. The subdivision of a cuboid stops when it contains at most N points, making it a leaf of the octree. N is chosen to yield the desired spatial resolution of the octree. To find a tetrahedron close to a desired source position $\mathbf{x}$, the octree is searched for the leaf cuboid enclosing $\mathbf{x}$. A tetrahedron with a vertex contained in that leaf cuboid lies close to $\mathbf{x}$, and can be used as a starting point for the adjacency walk, thus reducing the iterations needed for the walk to terminate (see Fig. 5.4b, dark grey) as well as the running time of the selection algorithm (see Fig. 5.4c, O+AW).

5.3 Experimental evaluation

To evaluate the performance of the proposed HRTF interpolation framework, experiments are carried out on both modelled and measured data.

To illustrate the qualitative effect of various approaches to select a subset of HRTF measurements for interpolation and to calculate interpolation weights on the estimated HRTFs, interpolation is performed on modelled HRTF data. The HRTF data is obtained via the ILD model in Eq. (2.7), and sampled at regular intervals of azimuth and elevation. Four different approaches for subset selection and interpolation weight calculation are compared:

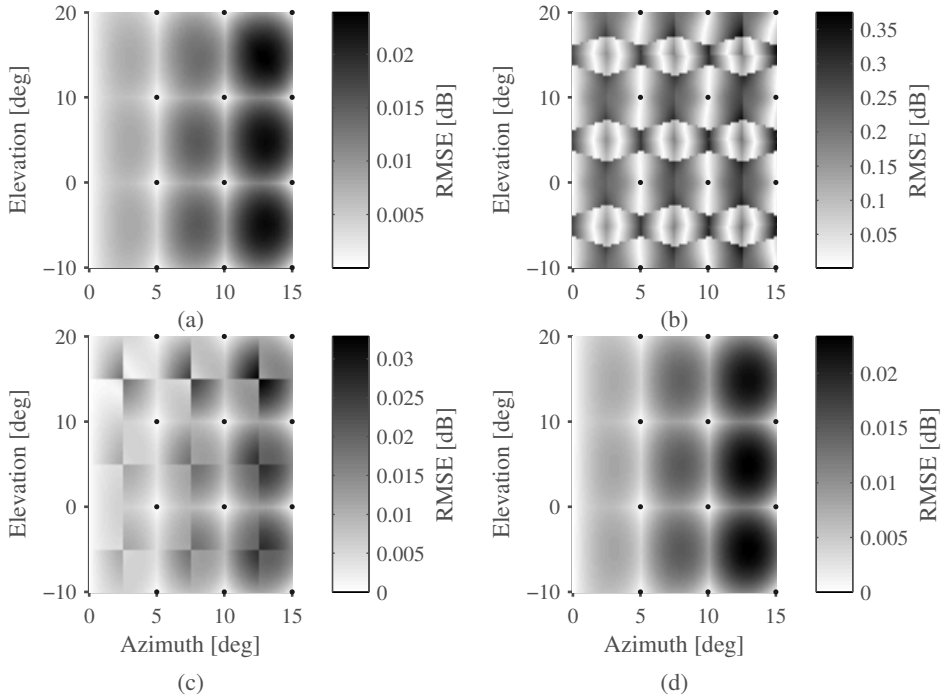

Figure 5.5. RMSE for interpolating a simple head-shadowing model sampled at regular azimuth and elevation intervals (dark dots) using (a) proposed method; (b) inverse distance weighting; (c) bilinear interpolation of three measurement points; (d) bilinear interpolation of four measurement points.

a) the proposed method (using triangulation and barycentric weights);

b) inverse distance weighting (Zotkin et al., 2004);

c) bilinear interpolation of three measurement points forming a triangle (Freeland et al., 2004);

d) bilinear interpolation of four measurement points forming a rectangle (Savioja et al., 1999).

For each approach, a subset of modelled HRTF data is selected and interpolated using the interpolation weights. The interpolation itself is performed the same way for all approaches, on the magnitude responses of the selected HRTFs.

Figure 5.5 illustrates the RMSE of the interpolation as a function of azimuth and elevation, for all four tested approaches. While the RMSE of the proposed method (Fig. 5.5a) and the bilinear interpolation of four measurement points (Fig. 5.5d) changes smoothly over the range of tested directions, both the inverse distance weighting (Fig. 5.5b) and the bilinear interpolation of three points (Fig. 5.5c) exhibit discontinuities. Note that the modelled data varies smoothly as a function of the lateral angle, hence the discontinuities visible in Fig. 5.5 are artefacts of the subset selection and interpolation weight calculation. Informal listening tests indicate that these discontinuities may be audible when rendering a moving virtual source.

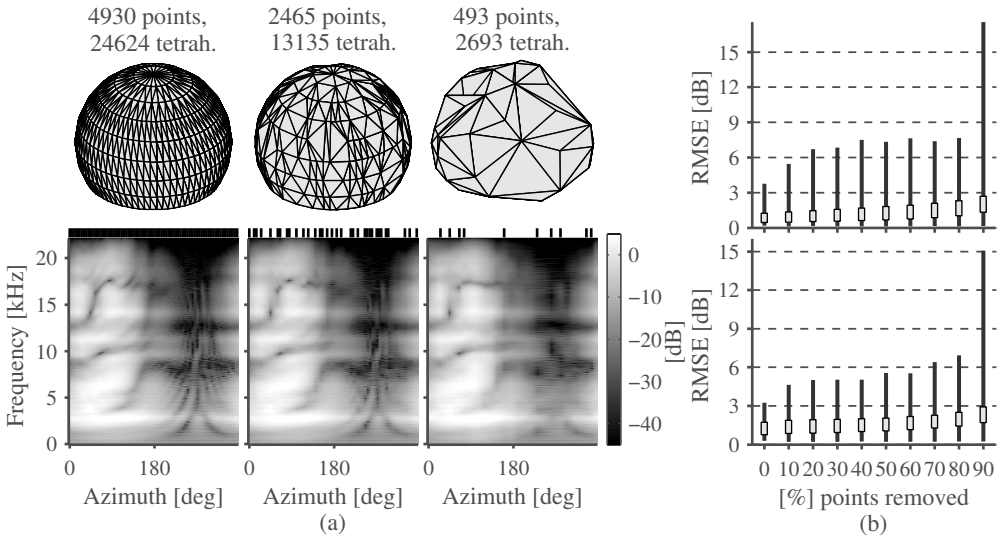

Figure 5.6. (a) Triangulation (top) and HRTF magnitude spectra (bottom) of the HRTF database by Yu et al. (2010), at 0 degrees elevation, 50 cm distance, using all measurement points, and with 50% and 90% of points randomly removed. Ticks mark magnitude spectra obtained directly from measured HRTFs (i.e., without interpolation). (b) RMSE for Qu et al. (2009) (top) and Yu et al. (2010) (bottom), as a function of the percentage of points removed; vertical lines extend from minimum to maximum, boxes from lower to upper quartile.

To evaluate the framework's performance for 3-D HRTF interpolation, experiments are carried out on datasets containing distance-dependent HRTF measurements: the PKU&IOA database (Qu et al., 2009), and the database by Yu et al. (2010). Figure 5.6a shows the triangulation and magnitude spectra for the database by Yu et al. (2010). To evaluate the effect of grid irregularities, datasets with measurements at irregular direction and distances were obtained by randomly removing measurement points from the HRTF datasets. This allows comparing interpolated and measured HRTFs at the removed positions as an objective performance measure. Examples of reduced datasets obtained from the database by Yu et al. (2010) are shown in Fig. 5.6a (top).

The HRTF interpolation is implemented as a linear combination of the measured HRTF magnitude spectra (Zotkin et al., 2004). The phase of the interpolated HRTF is not considered in the evaluation. It can be calculated using a spherical head model (Zotkin et al., 2004), and implemented, e.g., via Eq. (2.6), independent of distance (Brungart and Simpson, 2001). Interpolated magnitude spectra obtained via interpolation of reduced datasets are shown in Fig. 5.6a (bottom).

As an objective measure for the performance of the proposed interpolation framework, the RMSE between measured and interpolated HRTFs is calculated over third-octave bands with centre frequencies from 500 Hz to 16 kHz. The

RMSE for the two measured datasets as a function of the points removed to obtain the reduced datasets is shown in Fig. 5.6b. The RMSE is calculated for ten trials of random removal for each percentage, except for the 0 percent condition, which is obtained for 100 trials by removing a single random point from the measured dataset.

5.4 Discussion

A framework for HRTF interpolation in 2-D, with measurements on the surface of a sphere, and in 3-D, with measurements at various distances, is proposed. The main contributions of the proposed interpolation framework lie in the use of a standard triangulation method to efficiently group HRTF measurements into non-overlapping subsets, the use of a fast search algorithm to find a subset suitable for interpolation, and the use of triangular (in 2-D) or tetrahedral (in 3-D) interpolation using barycentric weights. An objective evaluation shows that the proposed framework is robust with respect to grid irregularities and produces HRTF estimates that change smoothly as a function of the source position, thus enabling the spatialisation of dynamic virtual sources in 2-D and in 3-D. A MATLAB® demonstration of the algorithm is available online[1].

[1] `http://www.mathworks.com/matlabcentral/fileexchange/43809` (Gamper, 2013)

6. Summary

This thesis studied various aspects related to the implementation of an audio augmented reality (AAR) system. A motion tracking algorithm was presented that relies on microphones embedded into a user-worn, binaural AAR headset to determine the user's position and orientation. The encoding of information into audible content via auditory display was studied, and results of a listening test indicating the effect of various display design parameters on user performance were presented. Finally, the rendering of spatialised virtual audio for the delivery of audible content as an overlay onto the real acoustic environment was investigated. As a result, a general framework for rendering sound in 3-D via head-related transfer function (HRTF) interpolation was proposed.

6.1 Main results

The main outcomes of this thesis can be summarised as follows:

- The head orientation tracking method proposed in Publication I successfully tracks user orientation based on speech signals recorded at binaural headset microphones. Unlike previously proposed methods, the method proposed here does not require anchor sources at known positions. By integrating distance estimates into the likelihood function of the tracking filter, the performance and robustness is substantially improved compared to a reference method.

- The position tracking method proposed in Publication II employs both user-worn binaural headset microphones and a reference microphone array to track user movement in a meeting scenario. Distance estimates are obtained both for the active speaker and the listeners, to derive an importance function for the tracking algorithm. The proposed importance function is shown to improve the accuracy and robustness of the tracking

algorithm. The motion tracking algorithm is applicable to other forms of user-worn microphones.

- Results from a listening test presented in Publication III suggest that with adequate stimulus onset asynchrony (SOA), users are able to detect a sample from distractors and estimate sample numerosity via auditory display. Samples automatically generated via text-to-speech synthesis led to similar performance as manually designed non-speech samples. The spatial quality of the samples did not affect performance, indicating that diotic or indeed monophonic playback may be sufficient in practical implementations. However, users were able to simultaneously detect and localise spatially presented samples.

- Previously proposed HRTF interpolation methods typically rely on ad hoc methods that rely on a particular measurement grid layout for selecting a subset suitable for interpolation. Experiments reported in Publication IV show that subset selection and interpolation weight calculation may introduce artefacts, regardless of the actual interpolation method used.

- The HRTF interpolation framework proposed in Publication V enables smooth and computationally efficient rendering of virtual sources. It relies on triangulation to group HRTF measurements into subsets for interpolation and barycentric coordinates to calculate interpolation weights. To the best of the author's knowledge, the proposed framework is the first that allows the direct interpolation of HRTF measurements taken at arbitrary directions and distances, in real time.

6.2 Future work

There are a number of possible future research directions:

- The motion tracking algorithms proposed in Publications I and II could be further improved by integrating dynamic models for the user movement and head rotation.

- The motion tracking algorithms could be integrated into a single framework, and the effect of replacing known user positions in the head orientation tracking algorithm (Publication I) with position estimates from the position tracking algorithm (Publication II) could be investigated.

- The findings of the listening test presented in Publication III could serve as the basis for the development and study of an AAR application conveying textual information to users outside laboratory settings.
- The HRTF interpolation framework presented in Publication IV and extended in Publication V could be further extended to allow the estimation of far-field HRTFs from near-field measurements.
- A perceptual evaluation of the proposed HRTF interpolation framework should be conducted to assess the fidelity of virtual sources rendered in the near field.
- The results presented in this thesis stem from simulation or user experiments in controlled laboratory settings. Further studies could investigate the performance of the proposed methods and the applicability of the results outside controlled environments.
- Finally, while this thesis studied various components of an AAR system in isolation, integrating the proposed methods and results into a single system would give further insights into the challenges and limitations arising from the combination of various components.

Bibliography

Ajanki, A., Billinghurst, M., Gamper, H., Järvenpää, T., Kandemir, M., Kaski, S., Koskela, M., Kurimo, M., Laaksonen, J., Puolamäki, K., Ruokolainen, T., and Tossavainen, T. (2011). An augmented reality interface to contextual information. *Virtual Reality* (**15**) (2), 161–173.

Albrecht, R. and Lokki, T. (07/2013). Adjusting the Perceived Distance of Virtual Speech Sources by Modifying Binaural Room Impulse Responses. In *Proc. Int. Conf. Auditory Display (ICAD)*. Lodz, Poland.

Albrecht, R., Lokki, T., and Savioja, L. (08/2011). A Mobile Augmented Reality Audio System with Binaural Microphones. In *Proc. of Interacting with Sound Workshop: Exploring Context-Aware, Local and Social Audio Applications*. Stockholm, Sweden, pp. 7–11.

Algazi, V. R., Duda, R. O., Thompson, D. M., and Avendano, C. (10/2001a). The CIPIC HRTF Database. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. New Paltz, NY, USA, pp. 99–102.

Algazi, V. R., Avendano, C., and Duda, R. O. (2001b). Elevation localization and head-related transfer function analysis at low frequencies. *J. Acoust. Soc. Am.* (**109**)3: 1110–1122.

Algazi, V. R., Duda, R. O., Duraiswami, R., Gumerov, N. A., and Tang, Z. (2002). Approximating the head-related transfer function using simple geometric models of the head and torso. *J. Acoust. Soc. Am.* (**112**)5: 2053–2064.

Anisetti, M., Ardagna, C., Bellandi, V., Damiani, E., and Reale, S. (2011). Map-Based Location and Tracking in Multipath Outdoor Mobile Networks. *IEEE Trans. on Wireless Communications* (**10**)3: 814–824.

Atal, B. S. and Schroeder, M. R. (02/1966). Apparent Sound Source Translator. Pat. US3236949: Filing date Nov 1962.

Aurenhammer, F. (1991). Voronoi diagrams – a survey of a fundamental geometric data structure. *ACM Comput. Surv.* (**23**)3: 09, 345–405.

Azuma, R., Baillot, Y., Behringer, R., Feiner, S., Julier, S., and MacIntyre, B. (2001). Recent advances in augmented reality. *Computer Graphics and Applications* (**21**)6: 34–47.

Azuma, R. T. (1997). A survey of augmented reality. *Presence: Teleoperators and Virtual Environments* (**6**)4: 355–385.

Azuma, R., Lee, J. W., Jiang, B., Park, J., You, S., and Neumann, U. (1999). Tracking in unprepared environments for augmented reality systems. *Computers & Graphics* (**23**)6: 787–793.

Batterman, J. M. and Walker, B. N. (07/2013). Auditory graphs need error bars: validating error-to-sound mappings and scalings. In *Proc. Int. Conf. Auditory Display (ICAD)*. Lodz, Poland.

Bederson, B. B. (1995). Audio augmented reality: a prototype automated tour guide. In *Proc. Conf. Human Factors in Computing Systems (CHI)*. Denver, Colorado, USA, pp. 210–211.

Begault, D. R., Lee, A. S., Wenzel, E. M., and Anderson, M. R. (02/2000). Direct Comparison of the Impact of Head Tracking, Reverberation, and Individualized Head-Related Transfer Functions on the Spatial Perception of a Virtual Speech Source. In *Audio Engineering Society Convention 108*. Paris, France.

Behringer, R., Chen, S., Sundareswaran, V., Wang, K., and Vassiliou, M. (1999). A novel interface for device diagnostics using speech recognition, augmented reality visualization, and 3D audio auralization. In *IEEE Int. Conf. on Multimedia Computing and Systems*. Florence, Italy, pp. 427–432.

Beracoechea, J. A., Torres-Guijarro, S., García, L., Casajús-Quirós, F. J., and Ortiz, L. (2008). Subjective Intelligibility Evaluation in Multiple-Talker Situation for Virtual Acoustic Opening-Based Audio Environments. *J. Audio Eng. Soc.* (**56**)5: 339–356.

Bernstein, J. G. W. and Grant, K. W. (2009). Auditory and auditory-visual intelligibility of speech in fluctuating maskers for normal-hearing and hearing-impaired listeners. *J. Acoust. Soc. Am.* (**125**)5: 3358–3372.

Blattner, M. M., Sumikawa, D. A., and Greenberg, R. M. (1989). Earcons and icons: their structure and common design principles. *Hum.-Comput. Interact.* (**4**)1: 11–44.

Blauert, J. (1996). *Spatial Hearing - Revised Edition: The Psychophysics of Human Sound Localization*. Cambridge, MA, USA: The MIT Press.

Bloom, P. J. (1977). Creating Source Elevation Illusions by Spectral Manipulation. *J. Audio Eng. Soc.* (**25**)9: 560–565.

Bolaños, J. G. and Pulkki, V. (10/2012). HRIR Database with Measured Actual Source Direction Data. In *Proc. Conv. Audio Eng. Soc.* 133. San Francisco, USA.

Bonebright, T. and Nees, M. (2009). Most earcons do not interfere with spoken passage comprehension. *Applied Cognitive Psychology* (**23**)3: 431–445.

Breebaart, J. (2013). Effect of perceptually irrelevant variance in head-related transfer functions on principal component analysis. *J. Acoust. Soc. Am.* (**133**)1: EL1–EL6.

Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, USA: The MIT Press.

Brewster, S. A. (2002). Overcoming the Lack of Screen Space on Mobile Computers. *Personal Ubiquitous Comput.* (**6**)3: 188–205.

Brewster, S. A., Raty, V.-P., and Kortekangas, A. (1995a). *Representing Complex Hierarchies with Earcons*. Tech. rep. ERCIM.

Brewster, S. A., Wright, P. C., and Edwards, A. D. N. (1993). An evaluation of earcons for use in auditory human-computer interfaces. In *Proc. Conf. Human Factors in Computing Systems (CHI)*. New York, NY, USA, pp. 222–227.

Brewster, S. A., Wright, P. C., and Edwards, A. D. N. (1995b). Experimentally Derived Guidelines for the Creation of Earcons. In *Proc. of BCS HCI*. Cambridge, UK, pp. 155–159.

Bronkhorst, A. W. (2000). The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker condition. *Acoustica* (**86**) 117–128.

Bronkhorst, A. W. (2002). Modeling auditory distance perception in rooms. In *Proc. Forum Acusticum*. Sevilla, Spain.

Bronkhorst, A. W. and Houtgast, T. (1999). Auditory distance perception in rooms. *Nature* (**397**) 517–520.

Brown, L., Brewster, S. A., Ramloll, R., Yu, W., and Riedel, B. (2002). Browsing Modes For Exploring Sonified Line Graphs. In *Proc. of BCS HCI*. London, UK, pp. 6–9.

Brungart, D. S., Ericson, M., and Simpson, B. D. (2002). Design considerations for improving the effectiveness of multitalker speech displays. In *Proc. Int. Conf. Auditory Display (ICAD)*. Kyoto, Japan.

Brungart, D. S. (02/2002). Near-field virtual audio displays. *Presence: Teleoper. Virtual Environ.* (**11**)1: 93–106.

Brungart, D. S. and Simpson, B. D. (2002). The effects of spatial separation in distance on the informational and energetic masking of a nearby speech signal. *J. Acoust. Soc. Am.* (**112**)2: 664–676.

Brungart, D. S., Simpson, B. D., Ericson, M. A., and Scott, K. R. (2001). Informational and energetic masking effects in the perception of multiple simultaneous talkers. *J. Acoust. Soc. Am.* (**110**)5: 2527–2538.

Brungart, D. and Simpson, B. (2001). Auditory localization of nearby sources in a virtual audio display. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. New Paltz, NY, USA, pp. 107–110.

Butz, A., Höllerer, T., Feiner, S., MacIntyre, B., and Beshers, C. (1999). Enveloping users and computers in a collaborative 3D augmented reality. In *Proc. of the 2nd IEEE and ACM Int. Workshop on Augmented Reality (IWAR)*. San Francisco, CA, USA, pp. 35–44.

Carlile, S., Jin, C. T., and Van Raad, V. (12/2000). Continuous virtual auditory space using HRTF interpolation: acoustic and psychophysical errors. In *Proc. IEEE Pacific Rim Conf. Multimedia*. Sydney, Australia, pp. 220–223.

Carty, B. and Lazzarini, V. (05/2009). Frequency-Domain Interpolation of Empirical HRTF Data. In *Proc. Audio Engineering Society Convention 126*. Munich, Germany.

Cherry, E. C. (1953). Some Experiments on the Recognition of Speech, with One and with Two Ears. *J. Acoust. Soc. Am.* (**25**)5: 975–979.

Cho, K., Nishiura, T., and Yamashita, Y. (2010). Robust speaker localization in a disturbance noise environment using a distributed microphone system. In *Int. Symposium on Chinese Spoken Language Processing*. Tainan, Taiwan, pp. 209–213.

Chung, J., Kim, N., Kim, J., and Park, C.-M. (2001). POSTRACK: a low cost real-time motion tracking system for VR application. In *Proc. Seventh Int. Conf. Virtual Systems and Multimedia*. Berkeley, CA, USA, pp. 383–392.

Cohen, M. M. and Massaro, D. W. (1990). Synthesis of visible speech. *Behavior Research Methods, Instruments, & Computers* (**22**)2: 260–263.

Cohen, M. and Wenzel, E. M. (1995). The design of multidimensional sound interfaces. In *Virtual environments and advanced interface design*. New York, NY, USA: Oxford University Press, Inc., pp. 291–346.

Cooper, D. H. and Bauck, J. L. (1989). Prospects for Transaural Recording. *J. Audio Eng. Soc.* (**37**)1: 3–19.

Dalenbäck, B.-I., Kleiner, M., and Svensson, P. (1996). Auralization, Virtually Everywhere. In *Proc. Audio Engineering Society Convention 100*. Copenhagen, Denmark.

Darwin, C. J. and Hukin, R. W. (2000). Effectiveness of spatial cues, prosody, and talker characteristics in selective attention. *J. Acoust. Soc. Am.* (**107**)2: 970–977.

Delorme, A. and Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics. *J. of Neuroscience Methods* (**139**) 9–21.

Dingler, T., Lindsay, J., and Walker, B. N. (2008). Learnabiltiy of Sound Cues for Environmental Features: Auditory Icons, Earcons, Spearcons, and Speech. In *Proc. Int. Conf. Auditory Display (ICAD)*. Paris, France.

DiVerdi, S. and Höllerer, T. (2007). GroundCam: A Tracking Modality for Mobile Mixed Reality. In *Proc. IEEE. Conf. Virtual Reality*, pp. 75–82.

Douc, R. and Cappé, O. (09/2005). Comparison of resampling schemes for particle filtering. In *Proc. Image and Signal Processing and Analysis (ISPA)*. Zagreb, Croatia, pp. 64–69.

Duda, R. O. (1997). Elevation dependence of the interaural transfer function. In *Binaural and Spatial Hearing in Real and Virtual Environments*. (Ed.) R. H. Gilkey and T. R. Anderson. Mahwah, USA: Lawrence Erlbaum Associates, pp. 49–75.

Duda, R. O., Avendano, C., and Algazi, V. R. (1999). An adaptable ellipsoidal head model for the interaural time difference. In *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*. Washington, DC, USA, pp. 965–968.

Duda, R. O. and Martens, W. L. (1998). Range dependence of the response of a spherical head model. *J. Acoust. Soc. Am.* (**104**)5: 3048–3058.

Dudoit, S., Shaffer, J. P., and Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statist. Sci.* (**18**)1: 71–103.

Duraiswami, R., Zotkin, D. N., and Gumerov, N. A. (2004). Interpolation and range extrapolation of HRTFs. In *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*. Montreal, Canada, pp. 45–48.

Eramudugolla, R., Irvine, D. R., McAnally, K. I., Martin, R. L., and Mattingley, J. B. (2005). Directed Attention Eliminates "Change Deafness" in Complex Auditory Scenes. *Current Biology* (**15**)12: 1108–1113.

Fallon, M. and Godsill, S. (2010). Acoustic Source Localization and Tracking Using Track Before Detect. *IEEE Trans. Audio, Speech, Language Processing* (**18**)6: 1228–1242.

Feiner, S., Macintyre, B., Höllerer, T., and Webster, A. (1997). A touring machine: Prototyping 3D mobile augmented reality systems for exploring the urban environment. *Personal Technologies* (**1**)4: 208–217.

Foursa, M. (2004). Real-time infrared tracking system for virtual environments. In *Proc. of the 2004 ACM SIGGRAPH international conference on Virtual Reality continuum and its applications in industry*. Singapore, pp. 427–430.

Freeland, F. P., Biscainho, L. W. P., and Diniz, P. S. R. (2004). Interpositional Transfer Function of 3D-Sound Generation. *J. Audio Eng. Soc.* (**52**)9: 915–930.

Freeland, F. P., Biscainho, L. W. P., and Diniz, P. S. R. (6/2002). Efficient HRTF Interpolation in 3D Moving Sound. In *Proc. Audio Engineering Society Conference 22*. Espoo, Finland.

Gamper, H. and Lokki, T. (2011). Spatialisation in audio augmented reality using finger snaps. In *Principles and Applications of Spatial Hearing*. (Ed.) Y. Suzuki, D. Brungart, and H. Kato. Singapore: World Scientific Publishing, pp. 383–392.

Gamper, H. (2013). *3-D HRTF interpolation*. `http://www.mathworks.com/matlabcentral/fileexchange/43809`.

Gardner, W. (1997). Head tracked 3-D audio using loudspeakers. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. New Paltz, NY, USA.

Gardner, W. G. and Martin, K. D. (1995). HRTF measurements of a KEMAR. *J. Acoust. Soc. Am.* (**97**)6: 3907–3908.

Garzonis, S., Jones, S., Jay, T., and O'Neill, E. (2009). Auditory icon and earcon mobile service notifications: intuitiveness, learnability, memorability and preference. In *CHI '09: Proc. of the 27th Int. Conf. on Human factors in computing systems*. New York, USA, pp. 1513–1522.

Gaver, W. W. (1986). Auditory icons: using sound in computer interfaces. *Hum.-Comput. Interact.* (**2**)2: 167–177.

Gelhard, O. and Grone, C. (2010). Earphone and headset. US 2010/0172532 A1: US Patent App. 12/637,014.

Genuit, K. (06/1987). Method and Apparatus for Simulating Outer Ear Free Field Transfer Function. US4672569: Filing date Mar 1985.

Gilkey, R. H. and Anderson, T. R., (Eds) (1997). *Binaural and Spatial Hearing in Real and Virtual Environments*. Mahwah, USA: Lawrence Erlbaum Associates.

Glumm, M. M., Marshak, W. P., Branscome, T. A., Mc.Wesler, M., Patton, D. J., and Mullins, L. L. (1998). *A Comparison of Soldier Performance Using*

Current Land Navigation Equipment with Information Integrated on a Helmet-Mounted Display. Tech. rep. ARL-TR-1604. Army Research Laboratory, Aberdeen Proving Ground, MD.

Gumerov, N. A., O'Donovan, A. E., Duraiswami, R., and Zotkin, D. N. (2010). Computation of the head-related transfer function via the fast multipole accelerated boundary element method and its spherical harmonic representation. *J. Acoust. Soc. Am.* (**127**)1: 370–386.

Härmä, A., Jakka, J., Tikander, M., Karjalainen, M., Lokki, T., Hiipakka, J., and Lorho, G. (2004). Augmented reality audio for mobile and wearable appliances. *J. Audio Eng. Soc.* (**52**)6: 618–639.

Hartmann, W. M. and Wittenberg, A. (1996). On the externalization of sound images. *J. Acoust. Soc. Am.* (**99**)6: 3678–3688.

Hartung, K., Braasch, J., and Sterbing, S. J. (04/1999). Comparison of Different Methods for the Interpolation of Head-Related Transfer Functions. In *Proc. Audio Engineering Society Conference 16*. Helsinki, Finland, pp. 319–329.

Healey, C. G., Booth, K. S., and Enns, J. T. (1996). High-speed visual estimation using preattentive processing. *ACM Trans. Comput.-Hum. Interact.* (**3**)2: 107–135.

Hebrank, J. and Wright, D. (1974). Spectral cues used in the localization of sound sources on the median plane. *J. Acoust. Soc. Am.* (**56**)6: 1829–1834.

Henrysson, A. and Ollila, M. (2004). UMAR: Ubiquitous Mobile Augmented Reality. In *MUM '04: Proc. of the 3rd international conference on Mobile and ubiquitous multimedia*. College Park, Maryland, USA, pp. 41–45.

Hightower, J. and Borriello, G. (2001). Location systems for ubiquitous computing. *Computer* (**34**)8: 57–66.

Hoff, B. and Azuma, R. (10/2000). Autocalibration of an electronic compass in an outdoor augmented reality system. In *Proc. IEEE and ACM Int. Symp. Augmented Reality (ISAR)*. Munich, Germany, pp. 159–164.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian J. of Statistics* (**6**) 65–70.

Hornof, A. J., Zhang, Y., and Halverson, T. (2010). Knowing where and when to look in a time-critical multimodal dual task. In *Proc. of the 28th Int. Conf. on Human factors in computing systems*. New York, USA, pp. 2103–2112.

IEEE (11/2001). *IEEE standard for inertial sensor terminology*.

Ihlefeld, A. and Shinn-Cunningham, B. (2008a). Spatial release from energetic and informational masking in a divided speech identification task. *J. Acoust. Soc. Am.* (**123**)6: 4380–4392.

Ihlefeld, A. and Shinn-Cunningham, B. (2008b). Spatial release from energetic and informational masking in a selective speech identification task. *J. Acoust. Soc. Am.* (**123**)6: 4369–4379.

IRCAM (12/03/2013). *LISTEN HRTF database*. http://recherche.ircam.fr/equipes/salles/listen/. Last viewed March 12, 2013.

Jin, C., Leong, P., Leung, J., Corderoy, A., and Carlile, S. (2000). Enabling individualized virtual auditory space using morphological measurements. In *Proc. First IEEE Pacific-Rim Conf. on Multimedia*. Sydney, Australia, pp. 235–238.

Jin, C., Corderoy, A., Carlile, S., and van Schaik, A. (2004). Contrasting monaural and interaural spectral cues for human sound localization. *J. Acoust. Soc. Am.* (**115**)6: 3124–3141.

Jot, J.-M., Larcher, V., and Warusfel, O. (02/1995). Digital Signal Processing Issues in the Context of Binaural and Transaural Stereophony. In *Proc. 98th Conv. Audio Eng. Soc.* paper no. 3980. Paris, France.

Julesz, B. and Bergen, J. R. (1987). Textons, the fundamental elements in preattentive vision and perception of textures. In *Readings in computer vision: issues, problems, principles, and paradigms*. (Ed.) M. A. Fischler and O. Firschein. San Francisco, USA: Morgan Kaufmann Publishers Inc., pp. 243–256.

Julier, S., Baillot, Y., Lanzagorta, M., Brown, D., and Rosenblum, L. (10/2000). BARS: Battlefield Augmented Reality System. In *NATO Symposium on Information Processing Techniques for Military Systems*. Istanbul, Turkey.

Kajastila, R., Lokki, T., and Takala, T. (03/2007). Virtual Acoustic Spaces With Multiple Reverberation Enhancement Systems. In *Audio Engineering Society Conference 30*. Saariselkä, Finland.

Kan, A., Jin, C. T., and van Schaik, A. (2011). Psychoacoustic evaluation of different methods for creating individualized, headphone-presented virtual auditory space from B-format room impulse responses. In *Principles and Applications of Spatial Hearing*. (Ed.) Y. Suzuki, D. Brungart, and H. Kato. Singapore: World Scientific Publishing, pp. 303–313.

Kan, A., Pope, G., Jin, C., and van Schaik, A. (2004). Mobile spatial audio communication system. In *Proc. 10th Int. Conf. on Auditory Display (ICAD2004)*. Sydney, Australia.

Kan, A., Jin, C., and van Schaik, A. (2009). A psychophysical evaluation of near-field head-related transfer functions synthesized using a distance variation function. *J. Acoust. Soc. Am.* (**125**)4: 2233–2242.

Karshmer, A. I., Brawner, P., and Reiswig, G. (1994). An experimental sound-based hierarchical menu navigation system for visually handicapped use of graphical user interfaces. In *Proc. of the first annual ACM Conf. on Assistive technologies*. Marina Del Rey, California, United States, pp. 123–128.

Kato, H. and Billinghurst, M. (1999). Marker Tracking and HMD Calibration for a Video-based Augmented Reality Conferencing System. In *Proc. IEEE Int. Workshop on Augmented Reality*. San Francisco, CA, USA, pp. 85–94.

Katz, B. and Schönstein, D. (02/2013). Method for selecting perceptually optimal HRTF filters in a database according to morphological parameters. US2013/0046790A1: US Patent App. 13/640,729.

Katz, B. F. G. (2001). Boundary element method calculation of individual head-related transfer function. I. Rigid model calculation. *J. Acoust. Soc. Am.* (**110**)5: 2440–2448.

Katz, B. F. G. and Parseihian, G. (2012). Perceptually based head-related transfer function database optimization. *J. Acoust. Soc. Am.* (**131**)2: EL99–EL105.

Kidd, G. R., Watson, C. S., and Gygi, B. (2007). Individual differences in auditory abilities. *J. Acoust. Soc. Am.* (**122**)1: 418–435.

Kidd, G., Arbogast, T. L., Mason, C. R., and Gallun, F. J. (2005). The advantage of knowing where to listen. *J. Acoust. Soc. Am.* (**118**)6: 3804–3815.

Kidd, J. G., Mason, C. R., Best, V., and Marrone, N. (2010). Stimulus factors influencing spatial release from speech-on-speech masking. *J. Acoust. Soc. Am.* (**128**)4: 1965–1978.

Kim, S.-M. and Choi, W. (2005). On the externalization of virtual sound images in headphone reproduction: A Wiener filter approach. *J. Acoust. Soc. Am.* (**117**)6: 3657–3665.

Kleiner, M., Dalenbäck, B.-I., and Svensson, P. (1993). Auralization - An Overview. *J. Audio Eng. Soc.* (**41**)11: 861–875.

Knapp, C. and Carter, G. (1976). The generalized correlation method for estimation of time delay. *IEEE Trans. on Acoustics, Speech and Signal Processing* (**24**)4: 320–327.

Kolarik, A., Cirstea, S., and Pardhan, S. (2013). Discrimination of virtual auditory distance using level and direct-to-reverberant ratio cues. *J. Acoust. Soc. Am.* (**134**)5: 3395–3398.

Lacouture-Parodi, Y. and Habets, E. A. (2012). Crosstalk Cancellation System Using a Head Tracker Based on Interaural Time Differences. In *Proc. Int. Workshop on Acoustic Signal Enhancement*. Aachen, Germany, pp. 1–4.

Lacouture-Parodi, Y. and Habets, E. A. (05/2013). Application of particle filtering to an interaural time difference based head tracker for crosstalk cancellation. In *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*. Vancouver, Canada, pp. 291–295.

Laitinen, M.-V., Pihlajamäki, T., Lösler, S., and Pulkki, V. (4/2012). Influence of Resolution of Head Tracking in Synthesis of Binaural Audio. In *Audio Engineering Society Convention 132*. Budapest, Hungary.

Langendijk, E. H. A. and Bronkhorst, A. W. (2000). Fidelity of three-dimensional-sound reproduction using a virtual auditory display. *J. Acoust. Soc. Am.* (**107**)1: 528–537.

Lehmann, E. (2004). Particle filtering methods for acoustic source localisation and tracking. PhD thesis. Australian National University.

Lehmann, E., Johansson, A., and Nordholm, S. (2007). Modeling of motion dynamics and its influence on the performance of a particle filter for acoustic speaker tracking. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. New Paltz, NY, USA, pp. 98–101.

Lentz, T., Assenmacher, I., Vorländer, M., and Kuhlen, T. (2006). Precise Near-to-Head Acoustics with Binaural Synthesis. *J. Virtual Reality and Broadcasting* (**3**)2: 1–12.

Li, M., Kim, B., and Mourikis, A. I. (05/2013). Real-Time Motion Estimation on a Cellphone using Inertial Sensing and a Rolling-Shutter Camera. In *Proc. IEEE Int. Conf. on Robotics and Automation*. Karlsruhe, Germany, pp. 4697–4704.

Liao, L., Fox, D., and Kautz, H. (2007). Extracting Places and Activities from GPS Traces Using Hierarchical Conditional Random Fields. *Int. J. of Robotics Research* (**26**)1: 119–134.

Lindeman, R., Noma, H., and Barros, P. de (2007). Hear-Through and Mic-Through Augmented Reality: Using Bone Conduction to Display Spatialized Audio. In *Proc. Int. Symp. on Mixed and Augmented Reality (ISMAR)*. Nara, Japan, pp. 173–176.

Litovsky, R. Y. and Godar, S. P. (2010). Difference in precedence effect between children and adults signifies development of sound localization abilities in complex listening tasks. *J. Acoust. Soc. Am.* (**128**)4: 1979–1991.

Loomis, J. M., Golledge, R. G., and Klatzky, R. L. (1998). Navigation System for the Blind: Auditory Display Modes and Guidance. *Presence* (**7**) 193–203.

Luo, Y., Zotkin, D., Daume, H., and Duraiswami, R. (05/2013). Kernel regression for Head-Related Transfer Function interpolation and spectral extrema

extraction. In *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*. Vancouver, Canada, pp. 256–260.

MacDonald, J. A., Henry, P. P., and Letowski, T. R. (2006). Spatial audio through a bone conduction interface. *Int. J. of Audiology* (**45**)10: 595–599.

Macpherson, E. A. and Middlebrooks, J. C. (2002). Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited. *J. Acoust. Soc. Am.* (**111**)5: 2219–2236.

Martin, A., Jin, C., and van Schaik, A. (2009). Psychoacoustic Evaluation of Systems for Delivering Spatialized Augmented-Reality Audio. *J. Audio Eng. Soc* (**57**)12: 1016–1027.

McGookin, D. and Brewster, S. A. (2004a). Space, the final frontearcon: The identification of concurrently presented earcons in a synthetic spatialized auditory environment. In *Proc. Int. Conf. Auditory Display (ICAD)*. Sydney, Australia.

McGookin, D. K. and Brewster, S. A. (2004b). Understanding concurrent earcons: Applying auditory scene analysis principles to concurrent earcon recognition. *ACM Trans. Appl. Percept.* (**1**)2: 130–155.

McGookin, D. and Brewster, S. (2012). PULSE: The Design and Evaluation of an Auditory Display to Provide a Social Vibe. In *Proc. Conf. Human Factors in Computing Systems (CHI)*. Austin, Texas, USA, pp. 1263–1272.

Menzies, D. and Al-Akaidi, M. (2007). Nearfield binaural synthesis and ambisonics. *J. Acoust. Soc. Am.* (**121**)3: 1559–1563.

Michalski, R. and Grobelny, J. (2008). The role of colour preattentive processing in human–computer interaction task efficiency: A preliminary study. *Int. J. of Industrial Ergonomics* (**38**)3-4: 321–332.

Middlebrooks, J. C. and Green, D. M. (1991). Sound Localization by Human Listeners. *Annual Review of Psychology* (**42**)1: 135–159.

Middlebrooks, J. C., Makous, J. C., and Green, D. M. (1989). Directional sensitivity of sound-pressure levels in the human ear canal. *J. Acoust. Soc. Am.* (**86**)1: 89–108.

Milgram, P., Takemura, H., Utsumi, A., and Kishino, F. (11/1995). Augmented Reality: A Class of Displays on the Reality–Virtuality Continuum. In *Proc. of the SPIE Conf. on Telemanipulator and Telepresence Technologies*. Boston, Massachusetts, USA, pp. 282–292.

Minnaar, P., Plogsties, J., Olesen, S. K., Christensen, F., and Møller, H. (02/2000). The Interaural Time Difference in Binaural Synthesis. In *Proc. Audio Engineering Convention 108*. Paper number 5133. Paris, France.

Möhring, M., Lessig, C., and Bimber, O. (11/2004). Video see-through AR on consumer cell-phones. In *Proc. Int. Symp. on Mixed and Augmented Reality (ISMAR)*. Arlington, VA, USA, pp. 252–253.

Møller, H., Jensen, C. B., Hammershøi, D., and Sørensen, M. F. (1999). Evaluation of Artificial Heads in Listening Tests. *J. Audio Eng. Soc.* (**47**)3: 83–100.

Møller, H., Sørensen, M. F., Jensen, C. B., and Hammershøi, D. (1996). Binaural Technique: Do We Need Individual Recordings? *J. Audio Eng. Soc.* (**44**)6: 451–469.

Morimoto, M. and Aokata, H. (07/1984). Localization cues of sound sources in the upper hemisphere. *J. Acoust. Soc. Japan* (**5**)3: 165–173.

Mynatt, E. D., Back, M., Want, R., Baer, M., and Ellis, J. B. (1998). Designing audio aura. In *CHI '98: Proc. of the SIGCHI conference on Human factors in computing systems*. Los Angeles, CA, USA, pp. 566–573.

Nees, M. and Walker, B. (2009). Auditory Interfaces and Sonification. In *The Universal Access Handbook*. (Ed.) C. Stephanidis. New York, USA: L. Erlbaum Associates, pp. 507–522.

Nicholson, D. (2013). Augmented reality grows up. *Engineering Technology* (**8**)4: 32–35.

Novotny, P. M., Stoll, J. A., Vasilyev, N. V., Nido, P. J. del, Dupont, P. E., Zickler, T. E., and Howe, R. D. (2007). GPU based real-time instrument tracking with three-dimensional ultrasound. *Medical Image Analysis* (**11**)5: 458–464.

Olaizola, I. G., Martirena, I. B., and Kammann, T. D. (2006). MHP Oriented Interactive Augmented Reality System for Sports Broadcasting Environments. *J. Virtual Reality and Broadcasting* (**3**)13: 1–11.

Oppenheim, A. V., Schafer, R. W., and Buck, J. R. (1999). *Discrete-time Signal Processing (2nd Ed.)* Upper Saddle River, NJ, USA: Prentice-Hall, Inc.

Parseihian, G. and Katz, B. F. G. (2012). Rapid head-related transfer function adaptation using a virtual auditory environment. *J. Acoust. Soc. Am.* (**131**)4: 2948–2957.

Pavani, F. and Turatto, M. (2008). Change perception in complex auditory scenes. *Perception & Psychophysics* (**70**)4: 619–629.

Pentenrieder, K., Bade, C., Doil, F., and Meier, P. (11/2007). Augmented Reality-based factory planning - an application tailored to industrial needs. In *Proc. Int. Symp. on Mixed and Augmented Reality (ISMAR)*. Nara, Japan, pp. 31–42.

Peres, S. C., Best, V., Brock, D., Frauenberger, C., Hermann, T., Neuhoff, J. G., Valgerdaur, L., Shinn-Cunningham, B., and Stockman, T. (2008). Auditory Interfaces. In *HCI Beyond the GUI: Design for Haptic, Speech, Olfactory, and Other Nontraditional Interfaces*. (Ed.) D. Penrose and M. James. Waltham, USA: Morgan Kaufmann, pp. 147–195.

Pertilä, P., Korhonen, T., and Visa, A. (2008). Measurement combination for acoustic source localization in a room environment. *EURASIP J. on Audio, Speech, and Music Processing* (**2008**) 3.

Pulkki, V., Lokki, T., and Rocchesso, D. (03/2011). Spatial Effects. In *DAFX: Digital Audio Effects*. (Ed.) U. Zölzer. 2nd. Chichester, UK: Wiley, pp. 139–184.

Pulkki, V. (1997). Virtual Sound Source Positioning Using Vector Base Amplitude Panning. *J. Audio Eng. Soc.* (**45**)6: 456–466.

Qu, T., Xiao, Z., Gong, M., Huang, Y., Li, X., and Wu, X. (2009). Distance-Dependent Head-Related Transfer Functions Measured With High Spatial Resolution Using a Spark Gap. *IEEE Trans. Audio, Speech, Language Processing* (**17**)6: 1124–1132.

Queiroz, M. and Sousa, G. H. M. de (2011). Efficient Binaural Rendering of Moving Sound Sources Using HRTF Interpolation. *J. New Music Research* (**40**)3: 239–252.

Ramloll, R., Yu, W., Riedel, B., and Brewster, S. (2001). Using non-speech sounds to improve access to 2D tabular numerical information for visually impaired users. In *15th Annual Conf. of the British HCI Group*. London, UK, pp. 515–529.

Rämö, J. and Välimäki, V. (2012). Digital Augmented Reality Audio Headset. *J. Electrical and Computer Engineering* (**2012**) 13.

Rayleigh, L. (1907). On our perception of sound direction. *Philosophical Magazine* (**13**) 214–232.

Regenbrecht, H., Baratoff, G., and Wilke, W. (11/2005). Augmented reality projects in the automotive and aerospace industries. *IEEE Computer Graphics and Applications* (**25**)6: 48–56.

Reitmayr, G. and Drummond, T. (10/2006). Going out: robust model-based tracking for outdoor augmented reality. In *Proc. Int. Symp. on Mixed and Augmented Reality (ISMAR)*. Santa Barbara, CA, USA, pp. 109–118.

Romblom, D. and Cook, B. (10/2008). Near-Field Compensation for HRTF Processing. In *Proc. Audio Engineering Society Convention 125*. San Francisco, CA, USA.

Rossing, T. D. and Fletcher, N. H. (2004). *Principles of vibration and sound*. 2nd ed. Springer, New York.

Rozier, J., Karahalios, K., and Donath, J. (2000). Hear&There: An Augmented Reality System of Linked Audio. In *Proc. Int. Conf. Auditory Display (ICAD)*. Atlanta, Georgia, USA, pp. 63–67.

Sagi, D. and Julesz, B. (1985). "Where" and "what" in vision. *Science* (**228**)4704: 1217–1219.

Samet, H. (1989). Implementing ray tracing with octrees and neighbor finding. *Computers & Graphics* (**13**)4: 445–460.

Savioja, L., Huopaniemi, J., Lokki, T., and Väänänen, R. (1999). Creating Interactive Virtual Acoustic Environments. *J. Audio Eng. Soc.* (**47**)9: 675–705.

Sawhney, N. and Schmandt, C. (2000). Nomadic radio: speech and audio interaction for contextual messaging in nomadic environments. *ACM Trans. Comput.-Hum. Interact.* (**7**)3: 353–383.

Schmalstieg, D., Langlotz, T., and Billinghurst, M. (2011). Augmented Reality 2.0. In *Virtual Realities: Dagstuhl Seminar 2008*. (Ed.) S. C. G. Brunnett and G. Welch. Springer, pp. 13–37.

Schönstein, D. and Katz, B. F. G. (08/2010). HRTF selection for binaural synthesis from a database using morphological parameters. In *Proc. Int. Congress on Acoustics*. Sydney, Australia.

Schwartz, O. and Gannot, S. (02/2014). Speaker Tracking Using Recursive EM Algorithms. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (**22**)2: 392–402.

Shilling, R. D. and Shinn-Cunningham, B. G. (2002). Virtual Auditory Displays. In *Handbook of Virtual Environments: Design, Implementation, and Applications*. (Ed.) K. S. Hale and K. M. Stanney. Mahwah, NJ: Lawrence Erlbaum Associates. Chap. 4, pp. 65–92.

Shinn-Cunningham, B. G., Santarelli, S., and Kopco, N. (2000). Tori of confusion: Binaural localization cues for sources within reach of a listener. *J. Acoust. Soc. Am.* (**107**)3: 1627–1636.

Shinn-Cunningham, B., Lehnert, H., Kramer, G., Wenzel, E., and Durlach, N. (1997). Auditory Displays. In *Binaural and Spatial Hearing in Real and Virtual Environments*. (Ed.) R. H. Gilkey and T. R. Anderson. Mahwah, USA: Lawrence Erlbaum Associates, pp. 611–663.

Sielhorst, T., Feuerstein, M., and Navab, N. (2008). Advanced Medical Displays: A Literature Review of Augmented Reality. *J. of Display Technology* (**4**)4: 451–467.

Slaney, M. (1998). A critique of pure audition. In *Computational auditory scene analysis*. Hillsdale, NJ, USA: L. Erlbaum Associates Inc., pp. 27–41.

Smith, J. O. (2007). *Mathematics of the Discrete Fourier Transform (DFT) with Audio Applications*. 2nd ed. `http://www.w3k.org/books/`: W3K Publishing.

Sohn, J., Kim, N. S., and Sung, W. (1999). A statistical model-based voice activity detection. *IEEE Signal Processing Letters* (**6**)1: 1–3.

Spagnol, S., Geronazzo, M., and Avanzini, F. (08/2012). Hearing distance: A low-cost model for near-field binaural effects. In *Proc. European Signal Processing Conference (EUSIPCO)*. Bucharest, Romania, pp. 2030–2034.

Spors, S. and Ahrens, J. (05/2011). Efficient range extrapolation of head-related impulse responses by wave field synthesis techniques. In *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*. Prague, Czech Republic, pp. 49–52.

Starner, T., Mann, S., Rhodes, B. J., Levine, J., Healey, J., Kirsch, D., Picard, R. W., and Pentland, A. (1997). Augmented Reality Through Wearable Computing. *Presence* (**6**)4: 386–398.

Sun, H., Yan, S., and Peter Svensson, U. (10/2009). Robust spherical microphone array beamforming with multi-beam-multi-null steering, and sidelobe control. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. New Paltz, NY, USA, pp. 113–116.

Sundareswara, R. and Schrater, P. (2003). Extensible point location algorithm. In *Proc. Int. Conf. Geometric Modeling and Graphics*. London, UK, pp. 84–89.

Sundareswaran, V., Wang, K., Chen, S., Behringer, R., McGee, J., Tam, C., and Zahorik, P. (2003). 3D audio augmented reality: implementation and experiments. In *Proc. Int. Symp. on Mixed and Augmented Reality (ISMAR)*. Tokyo, Japan, pp. 296–297.

Sutherland, I. E. (12/1968). A head-mounted three dimensional display. In *Proc. joint computer conference (AFIPS)*. San Francisco, CA, USA, pp. 757–764.

Takemoto, H., Mokhtari, P., Kato, H., Nishimura, R., and Iida, K. (2012). Mechanism for generating peaks and notches of head-related transfer functions in the median plane. *J. Acoust. Soc. Am.* (**132**)6: 3832–3841.

Talantzis, F. (2010). An Acoustic Source Localization and Tracking Framework Using Particle Filtering and Information Theory. *IEEE Trans. Audio, Speech, Language Processing* (**18**)7: 1806–1817.

Tappan, P. W. (10/1964). Proximaural Loudspeakers (-Nearphones-). In *The AES 16th Convention Preprints*. Paper number 358. New York, NY, USA.

Therneau, T. M., Atkinson, B., and Ripley, B. (2011). *rpart: Recursive Partitioning*. http://cran.r-project.org/package=rpart.

Tikander, M., Härmä, A., and Karjalainen, M. (2004). Acoustic Positioning and Head Tracking Based on Binaural Signals. In *Proc. 116th Audio Engineering Society Convention*. Berlin, Germany, pp. 1–10.

Tran, T., Letowski, T., and Abouchacra, K. (2000). Evaluation of Acoustic Beacon Characteristics for Navigation Tasks. *Ergonomics* (**43**)6: 807–827.

Treisman, A. (1986). Preattentive processing in vision. In *Papers from the second workshop Vol. 13 on Human and Machine Vision II*. (**3**) Montreal, Canada, pp. 313–334.

Välimäki, V., Parker, J., Savioja, L., Smith, J., and Abel, J. (2012). Fifty Years of Artificial Reverberation. *IEEE Trans. Audio, Speech, Language Processing* (**20**)5: 1421–1448.

Vargas, M. L. M. and Anderson, S. (2003). Combining speech and earcons to assist menu navigation. In *Proc. Int. Conf. Auditory Display (ICAD)*. Boston, USA, pp. 38–41.

Villegas, J. and Cohen, M. (2010). HRIR~: modulating range in headphone-reproduced spatial audio. In *Proc. ACM SIGGRAPH Conf. Virtual-Reality Continuum and its Applications in Industry*. VRCAI '10. Seoul, South Korea, pp. 89–94.

Walker, B. N. and Lindsay, J. (2005). Navigation performance in a virtual environment with bonephones. In *Proc. Int. Conf. Auditory Display (ICAD)*. Limerick, Ireland, pp. 260–263.

Walker, B. N., Nance, A., and Lindsay, J. (2006). Spearcons: speech-based earcons improve navigation performance in auditory menus. In *Proc. Int. Conf. Auditory Display (ICAD)*. London, UK, pp. 63–68.

Walker, B. N., Lindsay, J., Nance, A., Nakano, Y., Palladino, D. K., Dingler, T., and Jeon, M. (2013). Spearcons (Speech-Based Earcons) Improve Navigation Performance in Advanced Auditory Menus. *Human Factors: The Journal of the Human Factors and Ergonomics Society* (**55**)1: 157–182.

Walker, B. N. and Nees, M. A. (2011). Theory of sonification. In *The Sonification Handbook*. (Ed.) T. Hermann, A. Hunt, and J. G. Neuhoff. Berlin, Germany: Logos Publishing House. Chap. 2, pp. 9–39.

Wang, L., Yin, F., and Chen, Z. (2009). Head-related transfer function interpolation through multivariate polynomial fitting of principal component weights. *Acoustical Science and Technology* (**30**)6: 395–403.

Want, R., Hopper, A., Falcão, V., and Gibbons, J. (01/1992). The active badge location system. *ACM Trans. Inf. Syst.* (**10**)1: 91–102.

Ward, A., Jones, A., and Hopper, A. (1997). A new location technique for the active office. *IEEE Personal Communications* (**4**)5: 42–47.

Ward, D. B., Lehmann, E., and Williamson, R. (2003). Particle filtering algorithms for tracking an acoustic source in a reverberant environment. *IEEE Trans. Speech and Audio Processing* (**11**)6: 826–836.

Ware, C. (2012). *Information Visualization: Perception for Design*. 3rd edition. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Warusfel, O. and Eckel, G. (03/2004). LISTEN – Augmenting everyday environments through interactive soundscapes. In *IEEE Workshop on VR for public consumption*. Chicago, IL, USA.

Welch, G. and Foxlin, E. (2002). Motion tracking: no silver bullet, but a respectable arsenal. *Computer Graphics and Applications, IEEE* (**22**)6: 24–38.

Wenzel, E. M. (2001). Effect of increasing system latency on localization of virtual sounds with short and long duration. In *Proc. Int. Conf. Auditory Display (ICAD)*. Espoo, Finland, pp. 185–190.

Wenzel, E. M. (04/1999). Effect of Increasing System Latency on Localization of Virtual Sounds. In *Proc. Audio Engineering Society Conference 16*. Rovaniemi, Finland.

Wenzel, E. M., Arruda, M., Kistler, D. J., and Wightman, F. L. (1993). Localization using nonindividualized head-related transfer functions. *J. Acoust. Soc. Am.* (**94**)1: 111–123.

Wenzel, E. and Foster, S. (10/1993). Perceptual consequences of interpolating head-related transfer functions during spatial synthesis. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. New Paltz, NY, USA, pp. 102–105.

Wightman, F. L. and Kistler, D. J. (1989). Headphone simulation of free-field listening. I: Stimulus synthesis. *J. Acoust. Soc. Am.* (**85**)2: 858–867.

Wightman, F. L. and Kistler, D. J. (1997). Factors affecting the relative salience of sound localization cues. In *Binaural and Spatial Hearing in Real and Virtual Environments*. (Ed.) R. H. Gilkey and T. R. Anderson. Mahwah, USA: Lawrence Erlbaum Associates, pp. 1–23.

Wilson, J., Walker, B., Lindsay, J., Cambias, C., and Dellaert, F. (10/2007). SWAN: System for Wearable Audio Navigation. In *IEEE Int. Symp. Wearable Computers*. Boston, MA, USA, pp. 91–98.

Wu, K., Goh, S. T., and Khong, A. (05/2013). Speaker localization and tracking in the presence of sound interference by exploiting speech harmonicity.

In *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*. Vancouver, Canada, pp. 365–369.

Xiao, T. and Huo Liu, Q. (2003). Finite difference computation of head-related transfer function for human hearing. *J. Acoust. Soc. Am.* (**113**)5: 2434–2441.

Yairi, S., Iwaya, Y., and Suzuki, Y. (2008). Influence of Large System Latency of Virtual Auditory Display on Behavior of Head Movement in Sound Localization Task. *Acta Acustica united with Acustica* (**94**)6: 1016–1023.

Yost, W. A. (06/1993). Perceptual Models for Auditory Localization. In *Proc. Audio Engineering Society Conference 12*. Copenhagen, Denmark, pp. 155–168.

Yost, W. A. (1997). The Cocktail Party Problem: Forty Years Later. In *Binaural and Spatial Hearing in Real and Virtual Environments*. (Ed.) R. H. Gilkey and T. R. Anderson. Mahwah, USA: Lawrence Erlbaum Associates, pp. 329–347.

Yu, G.-Z., Xie, B.-S., and Rao, D. (10/2010). Characteristics of Near-Field Head-Related Transfer Function for KEMAR. In *Proc. Audio Engineering Society Conference 40*. Tokyo, Japan.

Yu, X., Pan, A., Tang, L.-A., Li, Z., and Han, J. (08/2011). Geo-Friends Recommendation in GPS-based Cyber-physical Social Network. In *Proc. Int. Conf. Advances in Social Networks Analysis and Mining (ASONAM)*. Kaohsiung, Taiwan, pp. 361–368.

Zahorik, P., Wightman, F., and Kistler, D. (10/1995). On the discriminability of virtual and real sound sources. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. New Paltz, NY, USA, pp. 76–79.

Zahorik, P., Brungart, D. S., and Bronkhorst, A. W. (2005). Auditory Distance Perception in Humans: A Summary of Past and Present Research. *Acta Acustica united with Acustica* (**91**) 409–420.

Zeimpekis, V., Giaglis, G. M., and Lekakos, G. (12/2002). A taxonomy of indoor and outdoor positioning techniques for mobile location services. *SIGecom Exch.* (**3**)4: 19–27.

Zhang, W., Zhang, M., Kennedy, R., and Abhayapala, T. (02/2012). On High-Resolution Head-Related Transfer Function Measurements: An Efficient Sampling Scheme. *IEEE Trans. Audio, Speech, Language Processing* (**20**)2: 575–584.

Zhong, X., Premkumar, A., and Wang, H. (2014). Multiple Wideband Acoustic Source Tracking in Three Dimensional Space Using a Distributed Acoustic Vector Sensor Array. *IEEE Sensors*, (in press).

Zhong, X.-L. and Xie, B.-S. (2009). Maximal azimuthal resolution needed in measurements of head-related transfer functions. *J. Acoust. Soc. Am.* (**125**)4: 2209–2220.

Zimmermann, A. and Lorenz, A. (2008). LISTEN: a user-adaptive audio-augmented museum guide. *User Modeling and User-Adapted Interaction* (**18**)5: 389–416.

Zotkin, D., Duraiswami, R., and Davis, L. (2004). Rendering localized spatial audio in a virtual auditory space. *IEEE Trans. Multimedia* (**6**)4: 553–564.

Zotkin, D., Duraiswami, R., and Gumerov, N. (2009). Regularized HRTF fitting using spherical harmonics. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. New Paltz, NY, USA, pp. 257–260.

Zotkin, D., Hwang, J., Duraiswami, R., and Davis, L. (10/2003). HRTF personalization using anthropometric measurements. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. New Paltz, NY, USA, pp. 157–160.

Aalto-DD 39/2014

ISBN 978-952-60-5621-0
ISBN 978-952-60-5622-7 (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